Library of
The
University of Nebraska

530.5 P56r
v.1-2
1929-30

PHYSICS

VOLUME 1 JULY, 1929 NUMBER 1

THE PHYSICAL REVIEW SUPPLEMENT

CONDUCTED BY THE AMERICAN PHYSICAL SOCIETY

UNIV. OF NEBR. AUG - 2 1929 LIBRARY

STUDIES IN CONTEMPORARY PHYSICS

BOARD OF EDITORS

JOHN T. TATE, University of Minnesota

A. H. COMPTON K. T. COMPTON K. K. DARROW
E. C. KEMBLE C. E. MENDENHALL D. L. WEBSTER

CONTENTS

Probable Values of the General Physical Constants . . . RAYMOND T. BIRGE 1
The Corpuscular Properties of Light . . . . . ARTHUR H. COMPTON 74
Statistical Theories of Matter, Radiation and Electricity . . KARL K. DARROW 90

PUBLISHED QUARTERLY

EDITORIAL OFFICES: MINNEAPOLIS

THE AMERICAN PHYSICAL SOCIETY

OFFICERS OF THE SOCIETY

President:	HENRY G. GALE	University of Chicago, Chicago, Ill.
Vice-President:	W. F. G. SWANN	Bartol Research Foundation, Philadelphia, Pa.
Secretary:	W. L. SEVERINGHAUS	Columbia University, New York City
Treasurer:	G. B. PEGRAM	Columbia University, New York City
Editor:	JOHN T. TATE	University of Minnesota, Minneapolis, Minn.
Local Secretary for the Pacific Coast:	L. B. LOEB	University of California, Berkeley, Calif.

THE PHYSICAL REVIEW SUPPLEMENT

STUDIES IN CONTEMPORARY PHYSICS

THE PHYSICAL REVIEW SUPPLEMENT is published quarterly under the auspices of the American Physical Society. Its purpose is to provide those interested in developments in physics with comprehensive, authoritative and timely discussions of the problems which are, at the present time, of especial interest and importance.

It is intended that in so far as possible these discussions shall be written in such form as to interest and to inform those who are not specialists in the subjects under discussion.

It is hoped that the studies appearing in the SUPPLEMENT will give to the specialist the benefit of the view point, in a broad sense, of other specialists; to the student, a background of critical knowledge and an adequate bibliography of source material; to those whose time is absorbed in the teaching of physics, a stimulating account of the forward movement of their science.

THE PHYSICAL REVIEW SUPPLEMENT is published quarterly in January, April, July and October by the American Physical Society at 450 Ahnaip St., Menasha, Wis. The volumes begin with the July number. The first volume began in July, 1929.

SUBSCRIPTION PRICE

To members of the American Physical Society or of the American Chemical Society	**per year**	**$3.00**
To all others	**per year**	**4.00**
Single copies		**1.25**

All correspondence should be addressed to THE PHYSICAL REVIEW SUPPLEMENT, University of Minnesota, Minneapolis, Minnesota.

Application has been made for entry as second class matter at the post office at Menasha, Wisconsin.

Volume 1. *July, 1929* *Number 1*

THE

PHYSICAL REVIEW SUPPLEMENT

PROBABLE VALUES OF THE GENERAL PHYSICAL CONSTANTS

(as of January 1, 1929)

By Raymond T. Birge
University of California, Berkeley

Table of Contents

			PAGE
Introduction			1
Section	A	Velocity of light, c	8
"	B	Gravitation constant, G	10
"	C	Relation of liter to cm^3	11
"	D	Normal mole volume of ideal gas, v_n	11
"	E	Relation of int. to abs. electrical units	14
"	F	Atomic weights, H, He, N, Ag, I, C, Ca	18
"	G	Normal atmosphere, A_n	26
"	H	Absolute temperature of ice-point, T_0	28
"	I	Mechanical and electrical equivalents of heat	30
"	J	Faraday, F	33
"	K	Electronic charge, e	36
"	L	Specific charge of electron, e/m	43
"	M	Planck constant, h	48
"	N	Tables	58
		Table a Principal constants and ratios	59
		Table b Additional quantities evaluated	60
		Table c Miscellaneous derived constants	62
"	O	Conclusions	67

Introduction

Some of the most important results of physical science are embodied, directly or indirectly, in the numerical magnitudes of various universal constants, and the accurate determination of such constants has engaged the time and labor of many of the world's most eminent scientists. Some of these constants can be evaluated by various methods. Each has been investigated by various persons, at various times, and each investigation normally produces a numerical result more or less different from that of any other investigation. Under such conditions there arises a general and continuous need for a searching examination of the *most probable* value of each important constant. The need is general since every physical scientist uses such constants.

287551

The need is continuous since the most probable value of to-day is not that of to-morrow, because of the never-ending progress of scientific research. These remarks appear to the writer so self-evident that the mere statement of them may be deemed superfluous. However, in spite of these facts, an investigation of the values of general constants in current use in the literature reveals a surprising lack of consistency, both in regard to the actually adopted values and to the origin of such values. This is probably due to the fact that it is almost impossible to find a critical study of the best values, sufficiently up-to-date to be really reliable, and sufficiently detailed to explain the inconsistencies found among older tables.

The situation is much better in the case of selected groups of constants. Thus the best value of the atomic weight of each element is determined annually by certain atomic weight committees, and the need of such a list of atomic weights is obvious to every chemist. There is certainly a similar need in the case of the even more important constants such as the velocity of light, the charge of the electron, the Planck constant h, etc. In attempting to respond to this need, the writer has become only too well aware of the intrinsic difficulties involved, but at the same time he has become increasingly convinced of the existence of the need itself. The present investigation was undertaken only at the express request of others, and the results given here should be considered more as a presentation of the situation than as a final solution of the problem. To obtain a satisfactory and thoroughly reliable judgment in such matters, there is required the unbiased cooperation of many persons situated in scientific laboratories throughout the world. In the preparation of this paper, I have endeavored to obtain such cooperation by means of an extensive correspondence, but because of a necessary limitation of time, such correspondence has been confined almost entirely to this country. In addition I have received valuable advice and suggestions from various persons on the campus of this University. To the many persons—too numerous to name individually—who have thus aided by voice or by letter in the preparation of this paper, I desire to express my sincere thanks. At the same time I wish to request a continuance of such cooperation, not only from those who have already aided, but from all readers of the present article. The character of such assistance, and its need, will be clearer after a discussion of the situation.

The decision as to the most probable value, at a particular time, of any given constant, necessarily demands a certain amount of judgment. It however demands, far more, a knowledge of the facts, and such knowledge is not always easy to obtain. Similarly, each investigator uses a certain amount of judgment in the selection of his data and in the final conclusions reached. Here also, apparently, it is the facts which are even more important. For example, there is scarcely an important constant which does not require, in its evaluation, a knowledge of the values of certain other constants. I call these "auxiliary constants." What value of each auxiliary constant did the investigator use? Is this still the best value, and if not, what effect will the required revision have on the final result? The writer has found, to his surprise,

that in many cases it is quite impossible to determine, from a printed article, what values of certain auxiliary constants actually were adopted. Often it is not possible to determine even the formula used in calculating the final result. Many approximations are made in actual calculations, and these are ordinarily justified by the precision—or better, lack of precision—of the investigation in question. Occasionally, however, an investigator makes an approximation which is not justified by the circumstances of the case, and which produces an appreciable change in the final result. The precise method of calculation employed also affects the result, and often the adopted method is capable of improvement.

In the preceding paragraph I have listed several sources of possible error in a final published value, quite distinct from the excellence of the experimental work itself. Whether it is worth while for the reviewer to attempt to improve a result by investigating such potential sources of error, depends entirely on the particular case considered. Thus if an investigator has spent many years collecting certain experimental data, and the reduction of these data, to give the final desired result, can be made in several weeks, by the most reliable analytic methods, while a less trustworthy method of calculation can be made in several days, it seems obvious which method of calculation should be employed. Yet the writer has found in the literature cases where an appreciable part of the accuracy obtained by years of the most skillful experimental work has been sacrified, naturally unintentionally, by approximate or ill-advised methods of computation. In such cases it has seemed imperative to make a recalculation of the results, and this has accordingly been done. In many cases no recalculation was possible, since the original data needed for such a calculation had not been published. In other cases it was impossible to determine even what method of calculation had actually been employed.

In the present paper the writer has, then, attempted to carry out the following general procedure—

(1) Each general constant has been determined from the available data, beginning with that constant whose value depends least on other constants. The value thus adopted has then been used *consistently* in the calculation of each succeeding constant for which it is an "auxiliary constant." No attempt has been made to compare the results of different investigators until these have been made properly comparable by the use of the same value of each auxiliary constant. Failure to make such revision, in the present paper, in certain cases, is due to lack of available information on the matter, and similar failure in the case of previous articles of this nature may well be due to a similar cause. This point I should like to emphasize above all others.

(2) Each constant has been calculated from the available data by the use, as far as possible, of formulas which involve no approximations. Various specific cases will appear in the later sections of approximations which have thus been eliminated, as well as cases in which, from lack of information, it has not been possible to make an intelligent revision.

(3) Each constant has been recalculated, whenever it seemed necessary, by analytic methods—usually by the method of least squares. The proper reduction of data is a very extensive subject, and is treated in detail in numerous text-books and articles. It seems however desirable to call specific attention to certain points.

In the present investigation I have endeavored to obtain not only the most probable value of a given constant, but also its "*probable error.*" Some estimate of the probable error is often quite as important as the constant itself, but published lists of constants contain, in general, no precise estimate of error. The term "probable error" is a technical one, applied properly only when the errors are distributed on a Gaussian error curve. It is defined as the numerical magnitude which the *true error is as likely to exceed as not.* This is a definite concept and as such furnishes an *objective* criterion of error. The writer strongly advocates its use in all cases. Expressions such as "the maximum error," "the limit of uncertainty," "the possible uncertainty," etc, are inherently vague and technically undefinable. It may be noted, in passing, that in some cases one cannot tell whether the published error is intended to represent a probable error or a limit of error.[1]

With a Gaussian error curve ($y = e^{-x^2}$), there is one chance in 4.5 that the true error is at least twice the probable error, one in 22 that it is at least three times the probable error, one in 142 that it is at least four times, etc. One can arbitrarily choose, as the limit of error, a quantity which the true error will exceed only once in a thousand times. This quantity, called the "huge error," is 4.9 times the probable error. One can however equally well adopt a ratio of one to the million, or one to the hundred, as defining the limit of error. One of the greatest difficulties in the present work has been the proper weighting of various results, one of which is stated in terms of "limit of error," another in terms of a "probable error," another possibly with the error undefined, and still another with no definite statement as to accuracy. This is quite aside from the reliability or correctness of the method used by each investigator in arriving at his stated error.

Now in order to evaluate the probable error it is necessary to use the method of least squares. One great objection, it appears to me, to certain methods which have been proposed as substitutes for least squares, is that they give no objective criterion for the error. These substitute methods often have the additional objection that they give no unique solution of the data. The investigator is permitted to combine the data in various different ways, each leading to a different result.[2] A third objection is that they are usually less reliable than least squares. Graphical methods combine all of these objections. They are never as accurate as an analytic method, and have no unique solution. Not only will each different person obtain in general, a different result

[1] German scientists often use the term "mean error." This is a quantity which, when multiplied by 0.6745, gives the probable error.

[2] This is quite aside from the question of *weighting*, which *does* require the judgment of the investigator and should be made in all cases.

from the same set of data, but the *same* person will obtain a different result, each time he makes such a graph of a given set of data. Graphical methods are invaluable for a preliminary survey of the results, especially in order to determine, if possible, the proper *functional* form to be used. They should, however, never be employed in the final accurate evaluation of important constants.[3]

The chief objection to least squares work is the time required. This objection is hardly valid for simple functions such as $y=a_{00}$, and $y=a_{01}+a_{11}x$, and as a matter of fact, the great majority of all calculations involved in this paper concern only these two simple functions. The least squares formula for the value of a_{01} and of a_{11} can be found in numerous texts. The corresponding formula for the *probable error* of these two quantities (r_{01} and r_{11}) is, however, practically never given. One can find the general formula applicable to any function, and the particular formula for r_{00}, but not the explicit formulas for r_{01} and r_{11}. Since the error formulas involve quantities occurring in the original formulas for a_{01} and a_{11}, I quote these also. If p is the weight of any observation,

$$a_{01}=\frac{(\Sigma px)(\Sigma pxy)-(\Sigma py)(\Sigma px^2)}{D}$$

$$a_{11}=\frac{(\Sigma px)(\Sigma py)-(\Sigma p)(\Sigma pxy)}{D}$$

where

$$D=(\Sigma px)^2-(\Sigma p)(\Sigma px^2)\,.$$

Then

$$r_{01}=0.6745\left[\frac{(\Sigma pv^2)(\Sigma px^2)}{(n-2)D}\right]^{1/2}$$

$$r_{11}=0.6745\left[\frac{(\Sigma pv^2)(\Sigma p)}{(n-2)D}\right]^{1/2}$$

where n is the number of observations, and v, the residual, is the difference between the observed y and its value as calculated from the values of a_{01} and a_{11} just given. The well known expression for the probable error of a weighted average is

[3] Because graphical methods have been used in the literature, in one or two such cases, the writer has made some objective experiments on this point. With a set of data which could be satisfactorily represented by the equation $y=a_0+a_1x$, it was found that the value of a_0 and also of a_1, as determined from a graphical solution, deviated *on the average* from the least squares values by just the probable error of the least squares solution. This result was obtained with the cooperation of a number of persons having considerable experience in the drawing of graphs. For real experts the graphical error can probably be made as small as one-half the least squares probable error. Now one of the most fundamental rules of computing is that the error introduced by the calculation itself shall be negligibly small, compared to the probable experimental error. A graphical solution obviously fails to satisfy this criterion in the case cited, and I believe that the result deduced is fairly typical of graphical solutions of simple functions.

$$r_{00}=0.6745\left[\frac{(\Sigma pv^2)}{(n-1)(\Sigma p)}\right]^{1/2}.$$

Before closing the discussion of methods of calculation it seems necessary to call attention to the fact that one must use some judgment in applying the method of least squares. Otherwise the results may well be absurd. Such a solution applies *only* to observations which are affected merely by accidental errors of observation. If a particular observation deviates too widely from a smooth curve,[4] it should be rejected before attempting to treat the data by least squares. If the data are so few in number that the residuals cannot be expected to fall on any symmetrical error curve, the results of least squares are relatively uncertain. Fortunately one can judge this uncertainty. Thus the "probable error of the probable error," or the "uncertainty of the probable error," is given by

$$ur=0.4769r/n^{1/2}$$

where r is the probable error, u its proportional uncertainty, and n the number of observations. This shows that with only two observations, the probable error has an even chance of being in error by more than 33.7 percent of itself,[5] while for nine observations this drops to 15.9 percent, and for 25 observations to 9.5 percent. These results lead naturally to a consideration of the number of figures to be retained in all experimental values. The standard rule of computing is that one should retain two doubtful figures and should express the probable error to two significant figures. Thus 126.944 ± 0.046. This rule is based on the preceding formula. If one wrote, for instance, 126.94 ± 0.05, the change in the probable error is nine percent of itself, and in the number also, in this case, nine percent of the probable error. These arbitrary changes are of the same magnitude as the uncertainty of the probable error, for 25 observations, and so are not justified.

In most experimental work, in addition to the probable error due to purely accidental errors, to which least squares properly applies, there are many other sources of error, constant or systematic. Some of these may be known, others may be quite unsuspected. The investigator endeavors to make some estimate of the magnitude of each such error, and in this case also, it seems greatly preferable to give what appears to be the *probable error*. If each error can be either plus or minus, the final probable error is to be obtained from the square root of the sum of the squares of all the probable errors, including the least squares accidental error.[6] If on the other hand the assumed "limit

[4] The criterion of "huge error" already mentioned is a convenient one for the rejection of an individual observation, where now the probable error is calculated for any one observation, instead of for the average. Other criteria have been proposed and can be found in texts on methods of calculation.

[5] In the case of only two unweighted observations of a constant quantity, the *mean* error of the average equals merely the common deviation of each observation from the average. It is however more in harmony with the idea of probable error to take this mean error as a measure of the probable error, and this is frequently done.

[6] If the various quantities concerned appear as factors, one uses the proportional error. If they appear as terms, one uses the absolute error.

of error" is given for each source of error, then obviously the final "limit of error" must be taken as the *arithmetic sum* of the component limits of error.[7] Since in many, if not most cases, the uncertain systematic errors are much larger than the least squares probable error, it is sufficient to use only one digit in stating the final probable error and only one doubtful figure in the actual result. This is common scientific practise. Fortunately, in the case of some of the most important general constants, the final error is, on the face of the evidence, due almost entirely to the accidental errors of the individual observations. In such cases one is fully justified in giving the error to two digits. The writer has accordingly stated the errors to one or to two digits according to circumstances Often two digits are used when the first numeral in the error is a 1, 2, or 3, but only one digit for higher numerals. In this connection it is to be noted that more significant figures should be carried in the calculations than it is desired to retain in the final result, in order that this final result may be unaffected by any error arising purely from the calculation. Failure to observe this rule is the cause of several apparent discrepancies in published data.

Before proceeding to the detailed discussion of each constant, attention should be directed to two important sources of information now available.

(1) The International Critical Tables (McGraw-Hill Co., 1926)[8] publish on page 17 of volume I, a list of nine so-called "Accepted Basic Constants," each with its "Uncertainty." On page 20 is given a list of 21 constants derived from these basic constants, and also certain other conventional and experimental constants. One of these conventional constants (the normal atmosphere) is actually an experimental constant, and this with the other experimental constants and the nine basic constants make up essentially the list of constants considered in the thirteen sections (A to M) of the present paper.

Although the ICT list of constants was published in 1926, it had been adopted in 1923, and since then much important work on nearly every constant has appeared. It was prepared with the aid of various scientific societies as well as of individuals. The values given are not claimed to be the best values available at that time, although obviously an attempt was made to obtain the best values. The chief weakness of this list of constants is the lack of any statement as to the origin of each value. By correspondence and in other ways the writer has obtained such information, and specific references to this are made in the various sections to follow.

(2) The Geiger and Scheel "Handbuch der Physik"[9] contains, on pages 487–517 of volume II, an article by F. Henning and W. Jaeger on "The General Physical Constants." On pages 516–517 there is a list of 52 constants, basic and derived, and on the preceding pages, a brief statement as

[7] All rules for combining errors involve approximations of the type $(1+b)^n = 1+nb$, where b is the error of the quantity 1.

[8] Throughout the present paper, this publication will be abbreviated to ICT.

[9] This publication will henceforth be denoted as HP.

to the theoretical and numerical basis of each value.[10] In their article many approximations and sources of inconsistency in the literature are pointed out, but with one or two exceptions no attempt is made to recalculate data in order to improve the published values. The writer has been greatly aided by this article, and without it and the ICT tables, it is doubtful if the present work would have been attempted.

The Henning and Jaeger article was written in 1926, and so contains more recent information than that in the ICT. Part of this is reflected in the fact that these two tables do not exactly agree on the value of a single basic constant, with the exception of the electronic charge e. Even since 1926, however, much new experimental material has appeared, and one result is that practically every constant adopted in the present paper differs more or less in value from that given in either of these two preceding lists. In fact in the case of the great majority of the constants considered in sections A to M, the finally adopted value is based primarily on work which has appeared since 1926. In the case of most of the constants, the situation is now much more satisfactory than it was a few years ago. Specific conclusions on these matters are given in section O.

The values adopted in this paper are based on data available to the writer, from published articles or private communications, on January 1, 1929. Reference to later work has been made in footnotes, and in practically all cases such recent work merely furnishes further confirmation of the adopted values. Because of the great amount of correlation of data, especially in connection with the numerous derived constants, it has not seemed feasible to make any changes in the adopted values of the basic constants, after the date just given. The present section, and the conclusions in section O, are being written in April 1929, and in the latter section especially the general situation is considered as it appeared at the date of writing.

A preliminary report on this work was made to the American Physical Society[11] in December 1928, and at that time a mimeographed table of constants was sent for suggestions to a number of persons. No changes have been made since then in the magnitudes of the constants printed in the abstract, but several of the probable errors have been revised, and this revision affects many of the derived constants.

Section A

The Velocity of Light in Vacuum (c)

The history of the experimental attempts to measure the velocity of light is a long and brilliant one. An accurate summary of all numerical results to 1927, in which many errors in the literature are corrected, has

[10] In other volumes of the HP there appear special articles on "h" and on "e/m", as mentioned in the appropriate sections of this paper.

[11] R. T. Birge, Phys. Rev. **33**, 265 (1929), abstract 6.

been given by de Bray.[1] A good recent account of the experimental methods for measuring c, as well as the numerical results, is that by Ladenburg.[2]

The latest and most accurate direct determination of the velocity of light is that by Michelson,[3] in 1921–26. In fact this investigation so far surpasses in accuracy any of the preceding that it alone need be considered. Michelson[4] lists eight results, obtained from eight sets of observations taken at different times with different mirrors. Each of these results is weighted and the weighted mean is 299796 km/sec. The usual least squares formula for the *probable* error of the weighted mean, I find, gives ± 2.6 km/sec. The *mean* error is 3.9 km/sec. Michelson writes 299796 ± 4 km/sec. and thus evidently uses the mean error of the weighted average. No other explicit statement about the error is made. When the various sets of results are collected under the five different mirrors used, the agreement is quite remarkable, all five results varying only from 299797 to 299795 with a mean of 299796 as before. Ladenburg[2] considers somewhat critically the possibility of various constant errors and concludes that all such errors are negligibly small. Aside from the possibility of unsuspected constant errors it is thus evident that the probable error is at most 2.6 km/sec. and may be much less. It seems however more conservative simply to adopt as the probable error the ± 4 km/sec given without comment by Michelson, and now being generally used. I accordingly adopt

$$c = (2.99796 \pm 0.00004) \times 10^{10} \text{ cm} \cdot \text{sec}^{-1}. \quad (1)$$

The velocity of short standing electric waves has also been measured and shown to agree approximately with the velocity of light, within the rather large limits of error. Probably the best work on this is by Mercier.[5] He shows that the formation of stationary electric waves on wires is a much more complicated phenomenon than had previously been realized. His measured velocity along a wire is corrected, by the use of various theoretical considerations, to give the velocity in free space. His final result is 299700 ± 30 km/sec for the velocity in free space, but because of uncertainties in the theory, the error may be (and according to Michelson's result actually is) much greater.

The velocity of electromagnetic waves may be obtained *indirectly* from the measured ratio of the electrostatic (es) to the electromagnetic (em) system of electrical units, according to the generally accepted electromagnetic theory of light. The best value of this ratio, which is here denoted by c', is undoubtedly that found by Rosa and Dorsey.[6] They obtained $c' = 2.9971 \times 10^{10}$ cm $\cdot$ sec^{-1}. This result is regularly quoted[7] as 2.9971 ± 0.0003.

[1] E. J. G. de Bray, Nature **120,** 602 (1927).

[2] R. Ladenburg, Hand. d. Exp. Physik **18,** 1–36 (1928).

[3] A. A. Michelson, Astrophys. J. **65,** 1 (1927).

[4] A. A. Michelson, loc. cit. (reference 3) Table VIII.

[5] J. Mercier, J.phys. radium **5,** 168 (1924).

[6] E. B. Rosa and N. E. Dorsey, Bur. Standards, Bull. **3,** 433–604 (1907). See also pp. 605–622 for the various possible methods.

[7] See, for instance, Bur. Standards, Circular No. 60 (1st ed.) 14 (1916) and Ladenburg.[2]

The authors themselves, however, state[8] that the *maximum uncertainty* is 1 in 10000, and Dr. Dorsey has written me that he believes the accuracy is even greater. Their final result is the average of a very large number of individual results, taken at different times, under varying conditions, and showing remarkable consistency. Rosa and Dorsey make in addition an exhaustive study of the sources of systematic error. It therefore seems to the writer that about one part in 30000 is a very conservative estimate for the *probable* error, giving $c' = 2.9971 \pm 0.0001$. This is an interesting conclusion for it indicates that the Rosa and Dorsey value of c' has been more accurate than any direct determination of the velocity of light, up to the time of Michelson's latest work.[9]

This result is however in terms of international electrical units.[10] The relation of these to the absolute electrical units is discussed in section E ahead. Henning and Jaeger[11] show that, to obtain the true ratio between the *es* and the *em* system, in absolute units, the result of Rosa and Dorsey must be multiplied by $p^{1/2}$, where one int. ohm $= p$ abs. ohm. According to section E, $p = 1.00051 \pm 0.00002$. This gives a corrected value of $c' = (2.9979 \pm 0.0001) \times 10^{10}$ cm $\cdot$ sec^{-1}. It is in beautiful agreement with Michelson's recent value of c, and has, on the above assumptions, a probable error only about 2.5 times as large. Hence we conclude that $c' = c$, within limits of error. A strict weighted average of c' and c would give 2.99795, but in view of the uncertainty in the relative probable error of c and c', it does not seem worth while to make this very slight change in Michelson's value.

Section B

The Newtonian Constant of Gravitation (G) and the Mean Density of the Earth

The HP[1] gives a table of seven determinations of G, ranging from 6.60 to 6.70×10^{-8} dyne $\cdot$ cm$^2 \cdot$ g^{-2}. After comparing the most probable values adopted previously by various persons, Henning and Jaeger adopt 6.65. In their list of previous critical reviews, they omit Poynting's value[2] of 6.66 ± 0.01. The ICT[3] adopt as one of their basic constants $G = 6.66 \pm 0.01$.

Since the publication of these reviews, Heyl[4] has made what is undoubtedly the most reliable determination of G. His final result is

[8] Rosa and Dorsey, loc. cit. (reference 6) page 601.

[9] This of course implies the assumption that c' should *theoretically* be equal to c.

[10] In giving the units of c' as cm $\cdot$ sec^{-1} we have implicitly assumed that the unit of resistance in the *em* system is one cm $\cdot$ sec^{-1}. This is strictly true for the absolute *em* unit, but is not strictly true for the international unit (see section E). Rosa and Dorsey (reference 6, page 601) write more correctly $c' = 2.9971 \times 10^{10}$ cm$^{1/2} \cdot$ sec$^{-1/2} \cdot$ int $\Omega^{-1/2}$.

[11] HP **2**, 507–508.

[1] Henning and Jaeger, HP **2**, 491.

[2] Poynting, article on "Gravitation," Encyl. Britannica, XI Edition.

[3] ICT. **2**, 17.

[4] P. R. Heyl, Proc. Nat. Acad. Sci. **13**, 601 (1927).

$$G = (6.664 \pm 0.002) \times 10^{-8} \text{ dyne} \cdot \text{cm}^2 \cdot \text{g}^{-2}$$

This result is adopted here. It is based on five separate determinations varying from 6.661 to 6.667.

Assuming $R = 6.371 \times 10^8$ cm as the mean radius of the earth, as given in the HP[1], and[5] $g_{45} = 980.616$ cm $\cdot$ sec^{-2}, one obtains[1] $G \cdot \delta(\text{earth}) = 36.797 \times 10^{-8}$ sec^{-2}, where δ(earth) is the *mean density* of the earth. From the HP result $G = 6.65$, one gets $\delta(\text{earth}) = 5.53$ g $\cdot$ cm^{-3}. With the new result $G = 6.664$, one gets

$$\delta(\text{earth}) = 5.522 \pm 0.002 \text{ g} \cdot \text{cm}^{-3}$$

Section C

Relation of the Liter to the Cubic Decimeter (1000 cm^3)

The liter is defined as the volume of a kilogram of air-free water at its maximum density. In other words, the maximum density of water is, by definition, one kg $\cdot$ l^{-1}. The kilogram is defined as the mass of the prototype kilogram preserved in Paris. This original prototype kilogram was intended to be the mass of a cubic decimeter (dm^3) of water, at maximum density, but later determinations have shown that there is a slight discrepancy. The various experimental results are discussed by Henning and Jaeger.[1] The mean of the best determinations[2] is 1 liter $= 1000.027$ cm^3, and this value has been accepted in all recent tables.[3] Henning and Jaeger give no probable error for the result, but one unit in the last place seems a not unreasonable assumption. Hence I adopt

$$1 \text{ liter} = 1000.027 \pm 0.001 \text{ cm}^3 = 1.000027 \pm 0.000001 \text{ dm}^3$$

The maximum density of water $\delta_m(H_2O)$ is accordingly

$$1/1.000027 = 0.999973 \pm 0.000001 \text{ kg} \cdot \text{dm}^{-3} \text{ or g} \cdot \text{cm}^{-3}$$

It should be noted in conclusion, that it is customary to define 1 cc as liter/1000, while 1 cm^3 = liter/1000.027.

Section D

The Normal Mole Volume of an Ideal Gas

(v_n cm^3 $\cdot$ mole^{-1}, or R_n liter $\cdot$ mole^{-1})

The normal mole volume of an ideal gas is the volume occupied by one gram mole of an ideal gas, at 0°C, under one normal atmosphere pressure.

[5] This conventionally accepted value of g_{45} is due to F. R. Helmert, Encycl. Math. Wiss. VI, 1B, 96 (1910). See HP **2**, 489.

[1] HP **2**, 491–2.

[2] Three of the best determinations give 1000.029, 1000.026, and 1000.027 respectively.

[3] ICT. **2**, 18, Smithsonian Tables (7th edition), HP, reference 1.

This quantity can theoretically be determined from any real gas, by making the correction necessary to reduce the real gas to an ideal gas. Actually, however, only oxygen is used for accurate determinations. Oxygen is chosen because its atomic weight is 16.000 by definition, and hence there is no error in the resulting value of v_n, due to error in the atomic weight. This is not the case for any other gas. Moreover, as a result of extensive investigations, the necessary correction to change oxygen to an ideal gas is now known with considerable accuracy.

The ICT[1] gives as one of its basic constants, $v_n = 22.4115 \times 10^3$ cm^3. The HP[2] gives $v_n = 22.414_5 \times 10^3$ cm^3 or $R_n = 22.413_9$ liters. The discrepancy must be due to the use of different values of $\delta_n(O_2)$, the normal density of oxygen, or of $(1-\alpha)$, the correction factor due to the deviation of oxygen from an ideal gas.[3] Thus

$$v_n = \frac{32(1-\alpha)}{\delta_n(O_2)} = \frac{32(1-\alpha)}{L_n(O_2)} 1000.027 = R_n(1000.027) \tag{1}$$

where v_n is the normal mole volume in cm^3, R_n the same in liters, $\delta_n(O_2)$ the normal density of O_2, in grams per cm^3, and $L_n(O_2)$ the normal density in grams per liter. All these values correspond to normal gravity ($g_n = 980.665$). It is however customary among chemists to express the experimental results in terms of g_{45} (980.616). Such values will be denoted by v, δ, L, and R. Thus

$$R = M(1-\alpha)/L \tag{2}$$

where M is the molecular weight.

[1] ICT **1**, 17.

[2] HP **2**, 494. See pp. 492–494 for a discussion of the factors involved in the determination of v_n.

[3] The most general definition of α is $(1/pv)\ d(pv)/d(p)$, (temp = constant), so that it measures the proportional change in pv, per unit change in pressure, and has the dimensions of pressure^{-1}. To make the numerical values more definite, it is customary to write $\alpha = [1/(pv)_1]\ d(pv)/d(p)$, where $(pv)_1$ refers to unit pressure. In investigations on normal density or normal mole volume, it is natural to choose one atmosphere as the unit of pressure. Henning and Heuse (Zeits. f Physik **5**, 285, 1921) use one meter of mercury as the unit of pressure, and denote α by κ_t (see section H). Since the numerical magnitude of α is directly proportional to the size of the unit of pressure, we have $\kappa_t = 100\alpha/76$. On the other hand, Henning (HP **9**, 528) uses the symbol κ_t, but states that p is measured in atmospheres.

Within limits of error, the isothermal pv has been shown to be a linear function of p, for the so-called permanent gases O_2, N_2, H_2, etc., so that for such substances α is independent of p but is a function of temperature, and is more properly written α_t. The linear extrapolation of pv to $p = 0$ gives then $(pv)_0 = (1-\alpha)\ (pv)_1$. Now in the limit $p = 0$, any gas becomes, by definition, an ideal gas. Hence $(pv)_0$ is the constant pv of an ideal gas, and $(1-\alpha)$ is the factor which converts the real $(pv)_1$ corresponding to unit pressure, into the ideal $(pv)_0$, both at some definite temperature. $(1-\alpha)$ is often denoted by $(1+\lambda)$, and $(1-\alpha)$ or $(1+\lambda)$ may be defined as the ratio $(pv)_0/(pv)_1$ (See reference 7, ahead). Frequently v is so chosen (in magnitude or unit) that $(pv)_1$ is unity. α(or κ_t) is then numerically (but *not* dimensionally) the *slope* of the pv isothermal (see HP **9**, 528 and 538).

Henning and Jaeger[4] give a full table of experimental values of L, taken from a critical summary by Moles.[5] The mean value is $L = 1.42892$ g·l^{-1}. They give also a table of values of $(1-\alpha)$, leading to the mean value $(1-\alpha) = 1.00092$. This table does not contain a more recent result by Baxter and Starkweather[6] ($L = 1.42901$), but this is included in the discussion by Henning and Jaeger, and raises the mean L to 1.42893. From this, and the value of $(1-\alpha)$ just quoted, the HP gets its value of v_n. The more recent values of $(1-\alpha)$, however, average 1.00086, and this, taken with the Baxter and Starkweather value of L, gives $v_n = 22.4119 \times 10^3$ cm^3, in close agreement with the ICT value.

Since the HP article was written, Moles[7] has critically discussed the most probable value of v_n. This article, as well as all recent experimental work is in turn discussed in the latest (8th) report of the German committee on atomic weights.[8] Moles obtains an average $L = 1.42892 \pm 0.00002$, but in getting it, he quotes[9] $L = 1.42892$ as a more recent value of Baxter and Starkweather.[10] Their true value is 1.42897, and the correct mean L is accordingly 1.42893 ± 0.00002, as pointed out by the committee.[8] Moles'[7] mean value of $(1-\alpha)$ is 1.00091 ± 0.00003, and these last two values give $R = 22.4148 \pm 0.0007$ liters. The discussion of the German committee, however, indicates that the recent results of Baxter and Starkweather[10] ($L = 1.42897$, and $1-\alpha = 1.00092$) are the most trustworthy. They give $R = 22.4144$ liters.

Very recently Baxter and Starkweather[11] have recalculated their 1926 data in a more logical manner and obtain $L = 1.428965$ grams per liter, $(1-\alpha) = 1.000927$. These values are adopted here. As just noted, Moles assumes 2×10^{-5} as the absolute probable error in L, and 3×10^{-5} as the absolute probable error in $(1-\alpha)$. The latter error seems to be a conservative estimate, but in the case of L, it seems better to take 3×10^{-5} as the error, in view of the difference between the adopted value and the older values. Hence we write

$$L = (1.428965 \pm 0.000030) \text{ gram·liter}^{-1} \quad (g = 980.616)$$
$$1-\alpha = 1.000927 \pm 0.000030.$$

This gives

$$R = 22.4146 \pm 0.0008 \text{ liter·mole}^{-1} \quad (g_{45} = 980.616)$$
$$R_n = 22.4135 \pm 0.0008 \text{ liter·mole}^{-1} \quad (g_n = 980.665)$$
$$v_n = (22.4141 \pm 0.0008) \times 10^3 \text{ cm}^3\text{·mole}^{-1} \quad (g_n = 980.665).$$

[4] HP **2**, 493.
[5] E. Moles, J. chim. phys. **19**, 100 (1921).
[6] G. P. Baxter and H. W. Starkweather, Proc. Nat. Acad. Sci. **10**, 476 (1924).
[7] E. Moles. Zeits. f. anorg. allgem. Chem. **167**, 40 (1927).
[8] Berichte, **61B**, 1 (1928).
[9] E. Moles, reference 7, page 46.
[10] G. P. Baxter and H. W. Starkweather, Proc. Nat. Acad. Sci. **12**, 699 (1926).
[11] G. P. Baxter and H. W. Starkweather, Proc. Nat. Acad. Sci. **14**, 57 (1928).

Section E

Ratio of International (int) to Absolute (abs) Electrical Units

For practical convenience, the ohm, ampere, and volt have been defined, by international agreement,[1] in terms of definite physical apparatus.[2]

The international ohm (int. ohm) is defined as the resistance offered to an unvarying electric current by a column of mercury at the temperature of melting ice, 14.4521 grams in mass, of a constant cross-sectional area and of a length of 106.300 centimeters.

The international ampere (int. amp.) is the unvarying electric current which, when passed through a solution of nitrate of silver, in accordance with certain specifications, deposits silver at the rate of 0.00111800 of a gram per second.

The international volt (int. volt) is the electrical pressure which, when steadily applied to a conductor the resistance of which is one int. ohm will produce a current of one int. amp.

These international units are to be compared with the corresponding absolute units, with which they were of course identical, within limits of experimental error, at the time of adoption in 1908. One abs. ohm $= 10^9$ *em* units of resistance, the *em* unit, under the assumption that permeability is dimensionless, being one cm $\cdot$ sec^{-1}. Measurements of the abs. ohm have been made in a variety of ways, but all methods necessarily involve the measurement of length and time. The abs. ampere is 10^{-1} *em* units, the *em* unit being one dyne$^{1/2}$, again with the assumption of dimensionless permeability. The older measurements of the abs. ampere were made usually with the tangent galvanometer, but the best measurements have more recently been made with the current balance, the force between a moving and a fixed coil carrying a current being compensated by a known weight (in dynes).

The definition of the int. amp. just given is the primary definition, and I shall follow the ICT in designating the int. amp. so defined, and all quantities involving it, by the symbol "(a)." Now let

$$1 \text{ int. ohm} = p \text{ abs. ohm} \tag{1}$$

$$1 \text{ int. amp. (a)} = q \text{ abs. amp.} \tag{2}$$

then

$$1 \text{ int. coul. (a)} = q \text{ abs. coul.} \tag{3}$$

$$1 \text{ int. volt (a)} = pq \text{ abs. volt} \tag{4}$$

$$1 \text{ int. joule (a)} = pq^2 \text{ abs. joule} \tag{5}$$

[1] Intern. Conference on elect. Units and Standards (London, Oct. 1908).

[2] For a complete account of electric units and standards, see Circular, Bureau of Standards No. 60 (First edition 1916, second edition 1920). For a short account see Smithsonian Tables (7th edition) pp. XXXVI to XLIV. For a very clear popular description of the principles involved in measuring electrical quantities in terms of absolute units, see F. E. Smith, Phys. Soc. London, Proc. **37**, 101 (1925).

$$1 \text{ int. henry} = p \text{ abs. henry} \tag{6}$$

$$1 \text{ int. gauss} = q \text{ abs. gauss.} \tag{7}$$

The int. ohm can be constructed as a definite laboratory standard. This is not true of the int. amp. (a). Hence the 1908 London conference[1] appointed a committee to determine the e.m.f. of the Weston normal cell, in terms of the int. ohm and int. amp. The final value adopted by the committee (Jan. 1, 1911) was 1.0183 int. volts, at 20°C, which, to avoid ambiguity, is written 1.01830. This is effectively a new definition of the int. volt and to distinguish it, if necessary, from the primary definition, I again follow the ICT in writing int. volt (v). Similarly all units involving the Weston normal cell will be designated by "(v)." Let

$$1 \text{ int. volt (v)} = r \text{ abs. volt} \tag{8}$$

as contrasted with Eq. (4). It is now possible to use the int. volt (v) and the int. ohm to obtain a new (subsidiary) definition of the int. amp. Thus

$$1 \text{ int. amp. (v)} = r/p \text{ abs. amp.} \tag{9}$$

as compared to Eq. (2). Finally, in many investigations, a so-called"semi-absolute" volt has been used. This is defined as the e.m.f. required to force one *abs.* amp. of current through one *int.* ohm resistance. Hence from Eq. (1)

$$1 \text{ semiabs. volt} = p \text{ abs. volt.} \tag{10}$$

From Eqs. (8) and (10) one obtains

$$1 \text{ int. volt (v)} = r/p \text{ semiabs. volt.} \tag{11}$$

We have now to consider the most probable value of p and of q, and the difference, if any, between r and pq (or between r/p and q). These questions are discussed by Henning and Jaeger in the HP[3], and they conclude,

$$q = 1, \qquad p = 1.0005_0 \qquad r = pq = 1.0005_0.$$

On the other hand, the ICT[4] gives

$$q = 0.99993, \qquad p = 1.00052, \qquad r = 1.00042, \qquad \text{while} \qquad pq = 1.00045.$$

Hence $r/p = 0.99990 \neq q$. The correct determination of the best values of p and q is a very technical and extremely involved matter. Unfortunately, as just seen, there is no exact agreement on the subject, even in the most authoritative compilations, and it is therefore necessary to consider the original sources. Part of the present disagreement in the values of p and q is due to the fact that there is no standard international unit of resistance or of voltage. Each national laboratory has its own standards which differ more or less among themselves, and also may change with time. The values of p and q finally adopted here represent, as well as possible, *mean* values

[3] HP **2**, 498–502.
[4] ICT **1**, 18 and 27.

both in respect to place and to time. Fortunately the accuracy of these quantities is so great that any possible error in the finally adopted values is entirely immaterial in its effect on the many constants derived later in this paper.

The numerical relation of the int. and abs. ohm (value of p) rests at the present time chiefly on two extensive investigations, one by Smith,[5] at the National Physical Laboratory (NPL) of England, and the other by by Grüneisen and Giebe,[6] at the German Reichsanstalt. Smith obtained $p = 1.00052 \pm 0.00004$, while Grüneisen and Giebe found $p = 1.00051$. The latter investigators estimate their own probable error, as well as that of Smith, as about 3 parts in 10^5. In 1925 a special committee was appointed at the NPL to make a general investigation of the relation of the int. and abs. electrical units. This work is not yet complete, although statements as to progress are given in the Reports of the NPL for 1925, 1926, and 1927. It was stated in 1925[7] that a comparison of various manganin resistances with mercury resistances indicates that the former have all increased in resistance by about 2.5 parts per 10^5, since 1912, or that the mercury standards (defining the int. ohm) are really smaller by this amount. The latter assumption would give $p = 1.000495$, in place of Smith's value of 1.00052. In a very recent investigation at the Reichsanstalt, Steinwehr and Schulze[8] evidently assume that the NPL 1925 standards are 2 parts in 10^5 less than the older 1912 standards, giving a mean value of p in exact agreement with the 1920 Reichsanstalt value. Their own experiments in 1928 agree with this same mean value to ± 1 in 10^5. Furthermore, various intercomparisons of standard resistances at the NPL[9] show that the German and American standards lie between the 1912 and 1925 NPL values. Hence it seems quite certain that the best value of p, at the present time, is 1.00051, while the probable error seems to be not more than 2 parts in 10^5.

The most probable value of q is more uncertain, in spite of extensive investigations on the subject. In the older work, the abs. amp., determined with either a current balance or a tangent galvanometer, was compared directly with the int. amp., as measured by a silver voltameter, thus evaluating q of Eq. (2). In other words, there was measured by means of a silver voltameter, operated under certain specifications, the amount of silver, in grams, deposited per sec. by a current of one abs. amp. This mass of silver was then compared with 0.00111800 grams, the defined amount deposited, *under the same conditions*, by one int. amp. per sec.

Such a procedure determines q unambiguously, but does not necessarily evaluate the electrochemical equivalent of silver (E_{Ag}) per abs. coul. As discussed in detail in section J, ahead, the electrochemical equivalent of a substance is the mass actually associated with unit charge, and is indepen-

[5] F. E. Smith, Phil. Trans. **214,** 27 (1914).
[6] E. Grüneisen and E. Giebe, Ann. Physik **63,** 179 (1920).
[7] National Phys. Lab. Reports p. 94 (1925).
[8] H. v. Steinwehr and A. Schulze, Ann. Physik **87,** 769 (1928).
[9] Nat. Phys. Lab. Reports p. 8 (1927).

dent of experimental imperfections, while the mass deposited in an electrolytic cell per unit charge,—the only quantity we can actually measure,—*is* subject to experimental imperfections. This distinction has no bearing on the value of q, so long as one accepts the official definition of the int. ampere. It concerns only the value of electrochemical equivalents and the resulting value of the Faraday, considered in section J. The matter is mentioned here because it is common practise among investigators to ignore the distinction, and instead of stating their real experimental values of q, to give the so-called corresponding E_{Ag} per abs. coul. ($=0.00111800/q$). The various experimental values of q, determined as explained above, are listed by Henning and Jaeger.[10] They are stated in terms of the assumed corresponding E_{Ag}, the actual values of which vary from 0.0011179 to 0.0011192 grams (i.e., q varying from 1.00009 to 0.99892).

In the later work (1906 to date) the current is measured in abs. amp., usually with a current balance, and this current is sent through an int. ohm resistance, using a Weston normal cell. From the known current in abs. amp. and the known resistance in int. ohms, one obtains the e.m.f. of the Weston cell in semi-absolute volts. By Eq. (11) the ratio of this result to the e.m.f. in int. volt (v), (1.01830 by definition), is r/p. Hence this later work evaluates only r/p, and not q.

The value of the e.m.f. of the Weston cell, in semi-abs. volts, the assumed corresponding electrochemical equivalent of silver per abs. coul., and the true resulting value of r/p, are listed by Henning and Jaeger.[11] Omitting a probably less accurate value by Guthe,[12] the remaining four values of r/p range from 1.00006 to 0.99989. Henning and Jaeger[11] give correctly as 1.01822 semi-abs. volts the Rosa, Dorsey and Miller[13] value of the e.m.f. of the Weston cell, but misquote and use in their averages the resulting E_{Ag} and r/p, giving 0.99995 for r/p in place of the true 0.99992 ($=1.01822/1.01830$). Using 0.99992, the unweighted average of the four investigations[14] is $r/p = 0.99995$. The Bureau of Standards[15] considers only (a) (c) and (d) of reference 14 and gives 0.99991 as the best average value of r/p. The ICT value (0.99990) is based[16] on (a) and (d) only, of reference 14. Henning and Jaeger[11] take the unweighted average of all four values of reference 14, and the writer has done the same, since there seems to be considerable difference of opinion among experts as to the relative weighting of these four values.

[10] HP **2,** 499.

[11] HP **2,** 500, Table 6.

[12] K. Guthe, Bur. Standards, Bull. **2,** 69 (1906).

[13] E. B. Rosa, N. E. Dorsey, and J. M. Miller, Bur. Standards, Bull. **8,** 269–393 (1912). See page 362.

[14] (a) Ayrton, Mather and Smith (NPL), 1908, $r/p=0.99989$, (b) Janet, Laporte and Jouaust (Lab. Central d'Electricité, Paris), 1908, $r/p=1.00006$, (c) Haga and Boerema (Univ. of Groningen, Holland), 1913, $r/p=0.99994$, (d) Rosa, Dorsey, and Miller (Bur. Standards, Washington), 1912, $r/p=0.99992$.

[15] Bur. Standards, Circular No. 60, page 38 (1916).

[16] Private communication from Dr. N. E. Dorsey.

In particular, it is very probable that (c) should be given a relatively lower weight, but the final average is fortunately not thereby changed.

The next question concerns the equality of r/p and q. Rosa, Vinal and McDaniel[17] determined the e.m.f. of the Weston cell as 1.01827 int. volt (a), by using a silver voltameter and an int. ohm resistance. Hence by Eqs.(4) and (8), knowing 1.01827 int. volt (a) = 1.01830 int. volt (v), $pq/r = 1.01830/1.01827 = 1.00003$. Hence $q = 1.00003\ r/p$. These investigators naturally assumed $r/p = 0.99992$, from reference 14(d). Hence $q = 0.99995$. This is the figure misquoted as r/p, by Henning and Jaeger.[11]

The above result indicates that q differs from r/p by 3 parts in 10,[5] and that, to agree with the primary int. units, the Weston cell should have been taken as 1.01827 int. volts. But at the Reichsanstalt,[18] the corresponding quantity was found, in 1908, to be 1.01834 int. volts, and in 1922, 1.01831. The average of these three results indicates that the accepted value of 1.01830 int. volts is correct, within limits of error. In other words, $q = r/p$, and one int. volt (a) = one int. volt (v). This agrees with the view of Henning and Jaeger.[19] The relative values of q and r/p adopted by the ICT[4] are based directly on the work of the Bureau of Standards[13,17].

Henning and Jaeger[19] consider that the variation from unity of either r/p or q is less than the experimental error, and under the circumatances think it more practical to assume $r/p = q = 1.0000$. It seems best, however, to accept the mean value of $r/p = 0.99995$, as determined in four different laboratories,[14] but to take the probable error as ± 0.00005. Assuming then no distinction between int. volt (v), and int. volt (a), we write equations (1) to (7)as

1 int. ohm $= p$ abs. ohm,	$p = 1.00051 \pm 0.00002$	(1′)
1 int. amp. $= q$ abs. amp.,	$q = 0.99995 \pm 0.00005$	(2′)
1 int. coul. $= q$ abs. coul.,		(3′)
1 int. volt $= pq$ abs. volt.	$pq = 1.00046 \pm 0.00005$	(4′)
1 int. joule $= pq^2$ abs. joule,	$pq^2 = 1.00041 \pm 0.00010$	(5′)
1 int. henry $= p$ abs. henry,		(6′)
1 int. gauss $= q$ abs. gauss,		(7′)

Section F

The Atomic Weights of Certain Elements (H, He, N, Ag, I, C, Ca).

In evaluating some of the general constants, it is necessary to use the atomic weights of various elements. Since it is only relative magnitudes of

[17] E. B. Rosa, G. W. Vinal, and A. S. McDaniel, Bur. Standards Bull. **10**, 475–536 (1914). See pp 477 and 486. See also reference 13, page 367.

[18] W. Jaeger and H. v. Steinwehr, Zeits. f. Instrumentenk. **28**, 327 and 353 (1908), and H. v. Steinwehr and A. Schulze, ibid, 42, 221 (1922).

[19] HP **2**, 501–2.

atomic weights which have physical significance, it follows that in the ultimate analysis, only ratios of atomic weights enter into our formulas for the general constants. All atomic weights are likewise derermined from ratios, but in general not directly from the particular ratios in which we are interested. Hence it is necessary to consider individual atomic weights, and for convenience, all those relevant to this paper are collected in the present section. The elements here discussed are hydrogen, helium, nitrogen, silver iodine, carbon, and calcium.

The present system of atomic weights is based on the purely arbitrary assumption that the weight of oxygen is 16 exactly. Any other assumed value for oxygen would be equally permissible, since as noted, atomic weights are entirely relative quantities. In choosing oxygen as a basis, it is tacitly assumed that this element has always the same atomic weight; in other words, that it has no isotopes. This however is seemingly not the case, for Giauque and Johnston[1] have very recently found an isotope of oxygen of atomic weight roughly 18, from an analysis of the atmospheric absorption bands of oxygen. The relative abundance of the two species of oxygen is not yet known, so that one cannot give accurately the new atomic weight of the "16" isotope, in terms of the standard mixture of the two isotopes taken as *exactly* 16. One practical result of this discovery is that the atomic weights determined recently by Aston,[2] from the mass spectrograph, need not be identical with those determined by chemical means, since Aston's atomic weights are based on the mass 16 isotope of oxygen considered as exactly 16, while the chemical atomic weights are based on the ordinary mixture of the two isotopes considered as exactly 16. As we shall see presently, Aston's values of the atomic weights of hydrogen, helium, nitrogen and iodine seem to agree with the chemical values within his *limit* of error(one part in ten thousand to one part in five thousand). If this is really the case, it means that, in ordinary oxygen, O_{16} is more than one thousand times as abundant as O_{18}.

The most detailed account of the very numerous investigations dealing with atomic weights, so far as I am aware, is that by F. W. Clarke.[3] This document of more than 400 large pages gives complete material on the subject, up to the date of publication. The method adopted by Clarke is to collect together all published determinations of a given atomic weight made by a given method, assign to each a probable error based entirely on the internal consistency of the data, and obtain a weighted average value with *its* probable error, using standard least squares formulas. No attention is given to the possibility of systematic errors, the author claiming that such errors will, in the end, more or less cancel out. His weighting of the results of different investigators is thus quite impersonal, being based wholly on internal consistency. He then collects together the determinations of a given atomic

[1] W. F. Giauque and H. L. Johnston, Nature **123,** 318 (1929), J. Am. Chem. Soc. **51,** 1436 (1929), and private communications. For more recent information see section O.

[2] F. W. Aston, Proc. Roy. Soc. **115A,** 487 (1927).

[3] F. W. Clarke "A Redetermination of Atomic Weights," 4th edition (1920). Printed as Part 3, Vol 16 of Memoirs, Nat. Acad. Sci.

weight, made by the various methods previously considered, and obtains a final weighted average and probable error. As might be expected, the average results from different methods often differ from each other by many times the probable error of each, thus clearly showing the presence of systematic errors in some or all of the methods employed. Whether such systematic errors do in fact cancel out in the long run may seriously be debated. The atomic weight committees, to be mentioned presently, seem more inclined to base the final value on what appears to be the most accurate determination by the most reliable method; in other words, on the determination in which it is most probable that systematic errors have been eliminated. The atomic weight committees also try to use ratios connected most directly with oxygen while Clarke considers all possible ratios.

Since the world war, there has been one report on the atomic weights of the elements by a so-called international committee.[4] Since 1925 the annual report on atomic weights by G. P. Baxter, for the American Chemical Society, has contained a table of atomic weights, for which he alone is responsible. Since 1918, a German committee on atomic weights has published annual reports. These various sources do not agree in many cases on the best value, and—what is more important for our purpose—do not give the probable error. There does not seem to be even a consistent practise in regard to the number of stated figures. Sometimes the last quoted figure is definitely doubtful, while in the great majority of cases it is certain. In the case of hydrogen the first five significant figures are certain, while only four figures are given by Baxter[5] and by the German committee.[6] The writer however is informed that the 1929 reports will give hydrogen as 1.0078. It is therefore necessary to investigate the probable values and probable errors, for the elements previously mentioned.

Hydrogen

Baxter[5] quotes Moles[7] as deciding, from a critical review of gas density measurements, that $H = 1.0078$. This value, as given by Moles[7] is actually taken from a previous article[8] where full details are given. Moles lists nine results lying in the narrow range 1.00766 to 1.00783, with a mean value of 1.00777 ± 0.00002, or a rounded figure of 1.0078. These nine results are given in Landolt-Börnstein[9], while Moles' article[8] is discussed in the 6th German report.[10] The final average represents the result of 223 different measurements by five different investigators, using four different methods, and seems to be the most reliable now available. It is quite possible that

[4] Second Report of the International Committee on Chemical Elements, J. Am. Chem. Soc. **47,** 597 (1925). This committee did not include German and Austrian scientists.

[5] G. P. Baxter, 34th annual report, J. Am. Chem. Soc. **50,** 603 (1928).

[6] German committee (8th report), Berichte, **61B,** 1–31 (1928).

[7] E. Moles, Gazz. chim. ital. **56,** 915 (1926).

[8] E. Moles, Zeits. f. physik. Chem. **115,** 61, and **117,** 157 (1925).

[9] Landolt-Börnstein, Physik. Chem. Tables, 5th ed. (E), p. 3 (1927).

[10] German committee (M. Bodenstein, etc.), Berichte, **59,** SII (A) (1926).

the probable error should be larger than that given by Moles, but even if it were many times larger, this would not affect any of the conclusions regarding the general constants. I accordingly adopt $H = 1.00777 \pm 0.00002$. Aston,[2] from positive ray analysis, obtains $H = 1.00778$ with a limit of error 0.00015. This last result is quoted by Baxter,[5] and is the only value for hydrogen mentioned in the latest German report.[6] As already pointed out, the recent discovery of an isotope of oxygen makes it now permissible to use Aston's value only as an indication of the relative abundance of O_{18} and O_{16}, and not as an atomic weight determination. It is in perfect agreement with the chemical value, which indicates a very low abundance of O_{18}.

Helium

In obtaining the best atomic weight of helium, one need consider only the most recent determinations. Baxter,[5] in his latest report, refers to an investigation by Baxter and Starkweather,[11] giving $L = 0.17846$ g·l^{-1} as the density of helium, for g_{45}. They also conclude that helium behaves like a perfect gas. Hence, using our adopted value[12] of R in the equation

$$R = M(1-\alpha)/L \tag{1}$$

we obtain

$$M = (22.4146 \pm 0.0008)(0.17846) = 4.00011 \tag{2}$$

as the atomic weight of helium. Baxter[5] then quotes unpublished work by R. B. Ellestad and himself, yielding $(1-\alpha) = 0.9995$. With this value, in Eq. (1), one obtains $M = 4.0021$. As far as chance variations are concerned, the probable error in L is one part in 20000 or less. The probable error in $(1-\alpha)$is not stated,[12a] but it is presumably at least one part in 10000. Hence the final probable error in M is at least ± 0.0004. Aston[2] obtained 4.00216 with a limit of error 0.0004, in perfect agreement with the revised Baxter and Starkweather value. As in the case of hydrogen, this indicates a small prevalence of O_{18} and encourages one to use Aston's values for other elements as true determinations of atomic weight.

The German committee[6] criticises the original Baxter and Starkweather value of $(1-\alpha) =$ unity, and points out that all previous work indicates that $(1-\alpha) = 0.9995$. It therefore also decides that He = 4.002, in agreement with Baxter.[5] It is evident that the true atomic weight of helium must be close to the value found by Aston. The chemical value is at present slightly less accurate than Aston's, and I shall accordingly adopt his value but his assumed error as the *probable* error, although such a procedure may well be

[11] G. P. Baxter and H. W. Starkweather, Proc. Nat. Acad. Sci. **12**, 20 (1926).

[12] See section D.

[12a] Just before this paper was to go to the printer, the writer received from Professor Baxter his individual values of $(1-\alpha)$ for He. There are four of these, averaging 0.99958, with a maximum variation of only 14 parts in 10^5. This new result would give He = 4.0018, with a final probable error of possibly 0.0003 or less. Since the *limit* of error of Aston's value is 0.0004, the chemical value is still the less accurate and no change is required in the final adopted value.

open to criticism, in view of the present situation regarding the oxygen isotopes. I accordingly write

$$He = 4.0022 \pm 0.0004.$$

Nitrogen

The atomic weight of nitrogen is not used directly in any of the following sections, but the error in this atomic weight produces practically the entire error in the atomic weight of silver. Since the great majority of the accepted atomic weights are derived more directly from silver than from oxygen, the atomic weight of silver is of the highest importance. It is therefore necessary to consider carefully the best value of N and its probable error.

The atomic weight of nitrogen can be obtained by direct comparison with oxygen, and also from density measurements, using the adopted value of R in Eq. (1). According to Clarke,[3] the final average of these two methods gives $N = 14.0076$. The atomic weight can be obtained indirectly in many ways. The results of all methods, including the two just mentioned, are summarized by Clarke[13] and give $N = 14.0081$. This was presumably the best value in 1920. At the present time it is generally agreed that, just as in the case of helium, the atomic weight of nitrogen can be determined most accurately from its density and deviation from a perfect gas, by the use of Eq. (1).

The various results have recently been discussed in full by Moles.[14] As explained in section D, Moles derives $R = 22.4148 \pm 0.0007$ g·l^{-1}, from data on oxygen, while we have adopted $R = 22.4146 \pm 0.0008$. For nitrogen, Moles quotes six values of L, varying from 1.25032 to 1.25071 g·l^{-1}, with a mean value 1.25048 ± 0.00006. Omitting the first two values, which he considers uncertain, the mean is 1.25046 ± 0.000045. This last value is adopted by Moles. Among the four values used in getting this average is 1.25049, obtained by Moles and Clavera.[15] This work represents probably the most accurate single investigation on nitrogen. The individual results are quoted in full by Baxter.[5] Moles and Clavera discuss carefully the various sources of systematic error, and conclude finally that $L = 1.25049 \pm 0.00003$. The chief reason for not adopting this as the best value of L is the lower result obtained by Baxter and Starkweather.[16] Their result is 1.25036, and they give no estimate of probable error, but remark that the result is slightly lower than the results of previous observers. It is therefore difficult to decide on the relative weighting of these various values. Moles and Clavera[15] give 1.25048 as the weighted average, including their own result. Moles,[14] as noted, gives 1.25046 ± 0.000045, based on the same available data. It appears to the writer that this result is not unreasonable, and it is adopted here.

[13] F. W. Clarke, reference 3, p. 118.
[14] E. Moles, Zeits. f. anorg. allgem. Chem. **167,** 40 (1927). See also section D.
[15] E. Moles and J. M. Clavera, Zeits. f. anorg. allgem. Chem. **167,** 49 (1927).
[16] G. P. Baxter and H. W. Starkweather, Proc. Nat. Acad. Sci. **12,** 703 (1926). See ibid **14,** 57(1928) for a better method of calculation which yields 1.25037.

The value of $(1-\alpha)$, or $(1+\lambda)$ as it is often denoted, is known with greater certainty for nitrogen than for oxygen.[12] The four best determinations, as quoted by Moles,[14] vary only from 1.00042 to 1.00046, with a mean of 1.000437 ± 0.000014. The value of $(1-\alpha)$ has been determined, indirectly, by Baxter and Starkweather,[16] as 1.00033. By another method of calculation, to which Moles[14] takes exception, they get 1.00051. Their recalculation[16] gives 1.00040. Moles himself concludes that, including also the Baxter and Starkweather value, $(1-\alpha) = 1.000424 \pm 0.000014$. Moles and Clavera[15] give 1.00043 as a final best value. Baxter and Starkweather give 1.00047 as the average of all work, excluding their own, but do not decide on a final average. I shall adopt $(1-\alpha) = 1.00043 \pm 0.00002$.

With the adopted values of R, L, and $(1-\alpha)$, we have in Eq. (1),

$$22.4146 \pm 0.0008 = \frac{M(1.00043 \pm 0.00002)}{1.25046 \pm 0.000045}. \qquad (3)$$

Hence $M = 28.01651$, or $M/2 = \mathrm{N} = 14.0083 \pm 0.0008$, the value adopted here. The error is larger than that adopted by Moles.[14] His own estimates of error for the three factors lead to one part in 20000 for the final probable error in N. He however states that the probable error is one part in 30000 (± 0.00045). Then in his summary he writes $\mathrm{N} = 14.0082 \pm 0.0002$, an error of only one part in 70000. This is presumably a misprint. Moles and Clavera,[15] with slightly different assumed values and errors for $(1-\alpha)$ and L, deduce[17] $\mathrm{N} = 14.0082 \pm 0.0004$.

In reviewing the literature on this subject, one is struck by the fact that, in the cases of both oxygen and nitrogen, the values finally adopted for both $(1-\alpha)$ and L tend to rise and fall together, so that the final result (value of R for oxygen, and the atomic weight for nitrogen) remains almost constant. A cursory survey of various critical summaries would thus indicate that the probable error in R and in the atomic weight of nitrogen is much smaller than that assumed here. Since the evaluation of $(1-\alpha)$ is quite independent from that of L, this seeming connection must be purely accidental. It is however quite possible that the error in N assumed above is a little high. Aston[2] obtains $\mathrm{N} = 14.008$, but his assumed accuracy in this case is only one part in 5000. It is to be recalled, in this connection, that Aston gives always the *limit* of error, and his *probable* error should be much smaller. Fortunately his values all agree beautifully with the chemical values, so that the decision as to his actual probable error may be left open.

Silver

As has already been stated, the best atomic weight of silver is at the present time directly dependent on that of nitrogen. A good summary of the situation is given by Moles and Clavera.[15] Of the many methods for obtaining the value of Ag, it is now generally agreed that the most accurate

[17] Written accidentally as ± 0.00004, but stated correctly as 3 parts in 10^5, and given in their summary correctly as ± 0.00042.

is that based on the reduction of $AgNO_3$ to Ag. Since O = 16.000, by definition, the sole error in Ag is due to that in N. Moreover, the proportional error in Ag is due to that in NO_3, and is therefore only about one-fourth the probable proportional error in N. The ratio $AgNO_3/Ag$ can be determined with great precision. Clarke[18] gives all the values of this ratio, up to 1920, the weighted average being 1.57479. Of these the most accurate, by far, is that by Richards and Forbes,[19] yielding also 1.57479. Since then a very elaborate investigation has been carried out by Hönigschmidt, Zintl and Thile,[20] which gives again exactly the same ratio. This latter work is discussed by Baxter[5] and by the German committee,[6] and is specifically accepted by the latter. There are twenty published values of the ratio which show a maximum range of only two parts in 150,000, and the probable error in the final mean is only about one part in 10^6. Hence, except for unsuspected systematic errors, the error in this ratio is quite negligible. With our adopted value of N, and with the above value of the ratio $AgNO_3/Ag = r$, one has

$$Ag = \frac{NO_3}{r-1} = \frac{62.0083 \pm 0.0008}{0.57479} = 107.8799 \pm 0.0014$$

If we had used the Moles[14] or the Moles and Clavera[15] value N = 14.0082, we would have obtained Ag = 107.8797. These results seem to indicate very strongly that the best value of Ag = 107.880. Hönigschmidt, Zintl and Thile[20] use the rounded figure N = 14.008, and therefore get Ag = 107.879.

The atomic weight of silver can be obtained in many other ways. Clarke[21] lists 43 different methods, yielding a final weighted average of 107.8804. It seems however clear at the present time that only the $AgNO_3/Ag$ ratio results need be considered, with a final real error in Ag due merely to that in N. It has already been pointed out that the error assumed here for N is larger than that assumed by others. It therefore does not seem unreasonable[21a] to adopt

$$Ag = 107.880 \pm 0.001$$

In order that the true result for Ag should be 0.001 more or less than this, N must be respectively 14.0089 or 14.0078.

Iodine

The atomic weight of iodine enters into the discussion of the value of the Faraday, in section J, and is therefore considered here. As in the case of silver, the atomic weight of iodine may be obtained in various different

[18] F. W. Clarke, reference 3, page 98.

[19] T. W. Richards and G. S. Forbes, J. Am. Chem. Soc. **29,** 808 (1907) and Zeits. f. anorg. Chem. **55,** 34 (1907).

[20] Hönigschmid, E. Zintl and P. Thilo, Zeits. f. anorg. allgem. Chem **163,** 65 (1927).

[21] F. W. Clarke, reference 3, page 116.

[21a] O. Hönigschmid and R. Sachtleben, Zeits. f. anorg. allgem. Chem. **178,** 1 (1929), have very recently been able to get a precise determination of silver in terms of oxygen, from the reduction of $Ba(ClO_4)_2$ and subsequent titration with silver, yielding finally Ag = 107.880 ± 0.001.

ways. Clarke[13] lists eight methods, with a mean result of 126.926. This result will, however, bear closer scrutiny. Of the eight methods, the most accurate is the direct determination of the I/Ag ratio, assuming the atomic weight of silver as known. All of the experimental results for this ratio are listed by Clarke.[22] There seems to be no more recent work on the subject. Among the values of this ratio, that obtained by Baxter,[23] in 1910, is the most reliable. His value of I/Ag is 1.176603, with an apparent probable error (from 13 determinations) of only 3 parts in 10^6. Clarke, following his usual custom, lists all determinations of this ratio. Now the four earlier results are all approximately 1.1753, while the later results run much higher. It therefore seems evident that these earlier results are vitiated by some systematic error. They are however quite self consistent, and so by Clarke's method are given a high weighting. Clarke's final weighted average is 1.176454. With the four earlier results eliminated, I obtain a new weighted average of 1.176549, in much closer agreement with Baxter's 1910 result. This revised ratio, combined with Ag = 107.880, gives I = 126.926, while Baxter's result gives 126.932. Using the revised average value for the I/Ag ratio with Clarke's results[13] for the other seven methods, I obtain a final weighted average of I = 126.932, in place of Clarke's value 126.926 previously quoted, and in exact agreement with Baxter's result. The International atomic weight committee,[4] in 1925, adopted I = 126.932, based entirely on Ag = 107.880 and Baxter's 1910 value of the I/Ag ratio. The German committee still retains[23a] the older international value 126.92. The above results seem to indicate quite definitely that I = 126.932, with a proportional error only slightly in excess of that for Ag. Hence I conclude $I = 126.932 \pm 0.002$.

Among the experimental values of the ratio I/Ag, as listed by Clarke[22] is that by Vinal and Bates,[24] from electrolysis. Their final value of the ratio is[25] 1.176235, leading to I = 126.892. This appears to be a relatively inaccurate determination of I, and is mentioned again in section J, in connection with the value of the Faraday. In conclusion it is of interest to note that Aston[2] gets I = 126.932, in exact agreement with our adopted value.

Carbon

The atomic weight of carbon can be determined directly from oxygen. The final result of all such determinations, as obtained by Clarke,[26] is 12.0000 ± 0.00026. This result (written 12.000) was accepted in 1925 by the International atomic weight committee,[4] and has since been used by Baxter.[5] There are however a number of other methods for obtaining C.

[22] F. W. Clarke, reference 3, p. 84.
[23] G. P. Baxter, J. Am. Chem. Soc. **32**, 1591 (1910).
[24] G. W. Vinal and S. J. Bates, Bur. Standards, Bull. **10**, 425 (1910).
[25] F. W. Clarke, reference 3, quotes 1.176260 (the reciprocal of 0.85015) while the true Vinal and Bates' value of Ag/I is 0.85017.
[23a] The 1929 report, Berichte **62**, 1 (1929), which has just appeared, gives 126.93. It also gives Ag = 107.880 in place of the older 107.88, and H = 1.0078, in place of the older 1.008. The other atomic weights of interest here are unchanged in value.
[26] F. W. Clarke, reference 3, p. 47.

Thirteen methods, including the above direct result, are listed by Clarke,[27] and give a weighted mean of 12.0025 ± 0.00019. Aston[2] finds $C = 12.0036$, with a limit of error 0.0012. It seems difficult to decide on the best chemical value. Clarke's printed probable errors as usual seem far smaller than the actual probable errors. We have now found that in the cases of H, He, N, and I, Aston's result agrees with the best chemical value well within his limit of accuracy, and this, it seems to the writer, furnishes sufficient justification for adopting a value of C which lies within Aston's stated limit of error. The mean of the final Clarke value (12.0025) and Aston's value (12.0036) is 12.003, and is adopted here. To be conservative, however, I adopt as the *probable* error essentially the value given by Aston as his *limit* of error, and write

$$C = 12.003 \pm 0.001.$$

Calcium

The atomic weight of calcium, as well as that of carbon, is needed for the evaluation of the grating space of calcite. The value of Ca which has been universally accepted since 1911 is 40.07. When this value was readopted in 1925 by the International committee, reference was made to the work of Richards and Hönigschmid.[28] These investigators precipitate calcium chloride by a solution of silver, and determine the amount of silver chloride produced. They assume $Ag = 107.88$ and $Cl = 35.457$. These are still the accepted atomic weights, that for Cl having recently been verified by Hönigschmid, Chan and Birckenbach.[29] The final result is $Ca = 40.075$, based on four determinations ranging from 40.085 to 40.070. Clarke[30] gives as the final result from all methods, $Ca = 40.0902$, but in obtaining this result he uses values for Cl, Br, and S not accepted at the present time. It seems probable that the Richards and Hönigschmid value of 40.075 is the best and is adopted here. The probable error seems to be about 0.005, and this figure is adopted, although it is admittedly very uncertain.

The atomic weights adopted in this section are then as follows:

$$H = 1.00777 \pm 0.00002 \qquad I = 126.932 \pm 0.002$$
$$He = 4.0022 \pm 0.0004 \qquad C = 12.003 \pm 0.001$$
$$N = 14.0083 \pm 0.0008 \qquad Ca = 40.075 \pm 0.005$$
$$Ag = 107.880 \pm 0.001$$

Section G

The Normal Atmosphere (A_n)

The normal atmosphere is defined as the pressure exerted by a column of mercury 76 cm high, of normal density (i.e., at 0°C, A_n), under normal

[27] F. W. Clarke, reference 3, p. 120.
[28] T. W. Richards and O. Hönigschmid, J. Am. Chem. Soc. **32,** 1577 (1910) and **33,** 28 (1911)
[29] O. Hönigschmid, S. B. Chan and L. Birckenbach, Zeits. f. anorg. allgem. Chem. **163,** 315 (1927).
[30] F. W. Clarke, reference 3, p. 144.

gravity. The very small change in the density of mercury, with change of pressure, permits an approximate value of A_n to be used, in determining the normal density. The normal atmosphere has been used in connection with certain preceding constants, but in such a way as not to involve its value in absolute units, as discussed in the present section.

The ICT[1] gives $A_n = 1.013250 \times 10^6$ dyne $\cdot$ cm^{-2}. This is based on the definition[2] of A_n as the pressure of a column of a liquid of density 13.5951 grams per cc, under normal gravity. The HP[3] gives $A_n = 1.01325_3 \times 10^6$, from the defining equation

$$A_n = H_n \cdot \rho_n(\mathrm{Hg}) \cdot \delta_m(\mathrm{H_2O}) \cdot g_n,$$

in which H_n = height of normal barometer = 76.000 cm,

ρ_n = normal specific gravity of Hg (at 0°C, A_n), referred to air-free water of maximum density,

$\delta_m(\mathrm{H_2O})$ = maximum density of water,

g_n = normal gravity[4] = 980.665 cm $\cdot$ sec^{-2}.

The ICT value of A_n follows directly from its definition, i.e.,

$$A_n = 76 \times 13.5951 \times 980.665 = 1.013250 \times 10^6 \text{ dyne} \cdot \text{ cm}^{-2}$$

Henning and Jaeger,[5] using the density of mercury in the definition, investigate the most probable value of ρ_n, and give a table of values taken from Scheel and Blankenstein,[6] which leads to an average value 13.59546. This table includes their own result of 13.59549. Henning and Jaeger then *adopt* $\rho_n = 13.5955$. The value of $\delta_m(\mathrm{H_2O})$ is[7] 0.999973 g $\cdot$ cm^{-3}. The product $\rho_n(\mathrm{Hg}) \cdot \delta_m(\mathrm{H_2O}) = D_n = 13.5955 \times 0.999973 = 13.595133$ g $\cdot$ cm^{-3}, agreeing with the ICT value of D_n to the six significant figures given by the ICT, but, with the use of seven figures, leading to $A_n = 1013253$, as given by the HP.[2] Hence the discrepancy in the seventh significant figure between the ICT and HP values of A_n seems to be due to carrying that result to seven figures, while the factors entering into it are given only to six.

I shall adopt as the most probable value of ρ_n, the figure calculated by Scheel and Blankenstein,[6] viz. 13.59546. This should have a probable error of a few units in the last digit. Then $D_n = 13.59546 \times 0.999973 = 13.59509$ g $\cdot$ cm^{-3}, and $A_n = 13.59509 \times 76 \times 980.665 = 1.013249 \times 10^6$ dyne $\cdot$ cm^{-2}. This in turn should have a probable error of not more than two or three units in the last digit. I adopt ± 0.000003.

The 45° atmosphere is obtained by the mere substitution of $g_{45}(980.616)$[8] for g_n. Hence

$$A_{45} = 13.59509 \times 76 \times 980.616 = (1.013199 \pm 0.000003) \times 10^6 \text{ dyne} \cdot \text{cm}^{-2}.$$

[1] ICT. **1**, 18.
[2] ICT. **1**, 34, definition of "Atmosphere."
[3] F. Henning and W. Jaeger, HP. **2**, 495.
[4] Adopted by the Fifth General Conference (1913), C. E. Guillaume, Trav. et Mem. du Bur. int. **16**, 114 (1913), See HP **2**, 490.
[5] HP. **2**, 490 and 494–496.
[6] K. Scheel and F. Blankenstein, Zeits. f. Physik **31**, 202 (1925). Given also in Landolt-Börnstein, 5th ed. (E) p. 12.
[7] See section C.
[8] See section B.

Note added in proof. It is evident from the preceding discussion that the definition of the normal atmosphere given by N. E. Dorsey in the ICT[2] is technically quite different from that given by Henning and Jaeger in the HP.[3] In particular, the ICT definition makes the normal atmosphere a conventional constant, with therefore no probable error. The writer had some correspondence on this matter with Dr. Dorsey, leading to the conclusion that the HP definition was correct. The adopted value is therefore based on this HP definition, as given in the first sentence of this section.

The writer unfortunately overlooked a recent article by G. K. Burgess[9] in which the "standard atmosphere" is defined as "the pressure due to a column of mercury 760 mm high, having a mass of 13.5951 $g \cdot cm^{-3}$, subject to a gravitational acceleration of 980.665 $cm \cdot sec^{-1}$, and is equal to 1,013,250 $dyne \cdot cm^{-2}$." It is thus a conventional constant, with no error. From direct correspondence with Dr. Burgess I now learn that this definition was adopted in 1927 by the International Commission of Weights and Measures. Fortunately this new definition makes no change in either the magnitude or error of any derived constant. It should be noted, in conclusion, that no temperature is specified in the definition and that the word "mercury" is technically superfluous. This seems to the writer very objectionable, since there is thus technically no simple method for reducing to standard atmospheres an actual barometer reading at an actual observed temperature. The HP definition, as used by the writer, therefore seems preferable, in spite of international agreement to the contrary.

Section H

The Absolute Temperature of the Ice-Point (T_0)

The generally accepted value of T_0 was, for many years, 273.09°K. This is based on Berthelot's analysis[1] of the data of Chappuis[2] giving $1/T_0 = \gamma = 366208 \times 10^{-8}$, and of Joule and Thomson for the porous plug experiment. The final average value was $\gamma = 36618 \times 10^{-7}$, or $T_0 = 273.09°$. The ICT[3] gives $T_0 = 273.1$ as one of its basic constants, a value apparently taken from Berthelot.

A most extensive series of observations on the volume and pressure coefficients (α and β) of certain gases has recently been carried out by Henning and Heuse,[4] at the Reichsanstalt. The value of γ was obtained by two different methods.[5] In the first it is assumed that the volume and pressure coefficients (α and β)[6] are linear functions of the pressure, so that their linear extrapolation to $p = 0$, gives γ ($= \alpha = \beta$ of an ideal gas $= 1/T_0$). In

[9] G. K. Burgess, Bur. Standards, J. Research **1**, 635 (1928).
[1] D. Berthelot, Trav. et Mem. du Bur. inter. **13**, 12 (1907).
[2] P. Chappuis, ibid. vols **6** and **13**.
[3] ICT. **1**, 17.
[4] (a) F. Henning, Zeits. f. Physik **5**, 264 (1921). (b) F. Henning and W. Heuse, ibid. **5**, 285 (1921). (c) W. Heuse, ibid. **37**, 157 (1926).
[5] See F. Henning, HP. **9**, 527–529, for a description of these methods.
[6] $(v_t/v_0)_p = 1 + \alpha t_p$, $(p_t/p_0)_v = 1 + \beta t_v$.

the second method it is assumed that the isothermal pv is a linear function of p. Choosing one meter of Hg as the unit of pressure and making pv unity for unit pressure, Henning and Heuse[4] denote the slope of the resulting pv curve by κ_t[7]. It can then be shown[8] that

$$\gamma = \beta + 3.73(\alpha - \beta) + 1.367 p_0 \kappa_0 \cdot 10^{-2}$$

where α and β are the coefficients for pressure p_0, taken from the data of the first method, and κ_0 refers to the 0°C isothermal.

From the first method they obtained for the gases He, H_2 and N_2, $\gamma \times 10^7 = 36600$, 36607, and 36606 respectively, or $T_0 = 273.224°$, 273.172° and 273.179°. The mean is $\gamma \times 10^8 = 366043$ or $T_0 = 273.190° \pm 0.015$.

From the second method they get for He (two determinations at slightly different p_0), H_2 and N_2, $\gamma \times 10^7 = 36598$, 36597, 36617, and 36604, respectively. The mean is 36604.0 or $T_0 = 273.194°$. They conclude that the best mean value of all the experiments is $\gamma \times 10^7 = 36604$. The reciprocal of this is $T_0 = 273.19°$. They write it, however, as 273.20°.

In the later article[4c] by Heuse, neon is used, and the above value of γ is confirmed. Because of a discrepancy between their value of κ_0, and that obtained by other observers, using high pressures,[9] a discrepancy which occurs in all the four gases used, Heuse writes the final probable error as $\pm 0.03°$, rather than the $\pm 0.01°$ which the above data seem to indicate.

The only other determination of T_0 of comparable accuracy is that by Roebuck,[10] using the Joule-Thomson effect in air.[11] This method requires a knowledge of α, the volume coefficient, as well as of the Joule-Thomson coefficient μ. Roebuck measured μ, but for α used older data, mainly by Chappuis. Henning and Jaeger[12] make note of this fact, and as the most probable value of T_0 they adopt merely the Henning and Heuse value 273.20° (which as previously noted should be 273.19°). Roebuck obtained three results, 273.18°, 273.16° and 273.12°, with an average of 273.15°. He gives a list of all previous determinations of T_0, and as the most probable value chooses 273.17°, a value lying midway between his own result and that of Henning and Heuse. He gives $\pm 0.02°$ as the probable error of the final average result. The writer feels that these two results (273.15° by Roebuck, and 273.19° by Henning and Heuse) are entitled to far more weight than any of the older work, but that the second result is probably the most accurate, being based on new determinations of α. Hence I adopt $T_0 = 273.18 \pm 0.03°$K. The probable error is that given by Henning and Jaeger,[12] and is a very conservative estimate. Roebuck's $\pm 0.02°$ may well be more reasonable.

[7] See footnote 3, section D, for the relation of κ_t to α, and the meaning of α (or κ_t) in terms of an ideal gas.

[8] F. Henning, Temperaturmessung, Braunschweig 1915, p. 68. See also references 5 and 4(b).

[9] See reference 4b for references.

[10] J. R. Roebuck, Proc. Am. Acad. Arts Sci. **60**, 537 (1925).

[11] See HP. **9**, 526–527, for the theory of this method.

[12] HP. **2**, 496.

Section I

The Mechanical Equivalent of Heat (J) and the Electrical Equivalent of Heat (J')

A detailed description of the methods which have been employed for the evaluation of J, and a critical discussion of the results, is given by Jaeger in the HP.[1] The value adopted by Henning and Jaeger in the HP[2] is one $cal_{15} = 4.184_2$ int. joule $= 4.186_3$ abs. joule. The ICT value[3] is one $cal_{15} = 4.185$ abs. joule. The cal_{15} is defined as the amount of thermal energy required to heat one gram of pure water from 14.5° to 15.5°C.

In the original work by Joule, mechanical energy was turned directly into thermal energy, and J was thus directly evaluated in abs. joules. In most of the modern work, however, electrical energy is turned directly into thermal energy, thus evaluating the *electrical* equivalent of heat (J', measured in int. joules). Since the relation between the int. joule and the abs. joule (10^7 ergs) is known with considerable precision,[4] the value of the mechanical equivalent of heat may be obtained from the observed value of the electrical equivalent.

The value of J adopted by the HP results from the work of Jaeger and Steinwehr.[5] These investigators determined J', for many different mean temperatures[6] lying between 4.75°C and 49.60°C. This is undoubtedly the most extensive and accurate work on J' now available. They list 67 results in all. These results, plotted against temperature, are represented as a parabolic function of t. The resulting least squares equation is

$$J'(\text{int. joule}) = 4.2047_7 - 1.768 \times 10^{-3} t + 2.644_7 \times 10^{-5} t^2, \qquad (1)$$

where t is measured from 0°C, on the thermodynamic scale. This equation gives 4.18420 int. joules for the 15° calorie. The HP writes this 4.184_2.

On examining their data, the writer finds that a parabola is not a sufficiently complex function to give a proper representation of the observations. Their residuals show pronounced trends, and unfortunately the largest trend is just in the vicinity of 15°C, where it is most important that the curve give a true most probable representation of the data. I have accordingly made a separate investigation of the best curve for the representation of their data. The detailed work will be presented in a separate paper, and only the final result is quoted here. This is in the form of a fourth degree polynomial

$$J' = 4.21040 - 2.78958 \times 10^{-3} t + 7.73723 \times 10^{-5} t^2 - 8.52567 \times 10^{-7} t^3 + 3.7540 \times 10^{-9} t^4 \qquad (2)$$

[1] W. Jaeger, HP. **9**, 476–495.
[2] HP. **2**, 497.
[3] ICT. **1**, 18.
[4] See section E.
[5] W. Jaeger and H. v. Steinwehr, Ann. Physik **64**, 305 (1921). See also HP **9**, 490–493.
[6] The temperature change in the calorimeter was usually about 1.4°C, for the six minute run used.

This gives $J'_{15}=4.18327$ int. joules, and is, I believe, the true most probable value resulting from the work of Jaeger and Steinwehr. It differs by 9.3×10^{-4} joules from their own deduced value.

These investigators consider with great care all of the possible errors, both systematic and accidental. They conclude that the electrical standards (normal resistance and normal Weston cell) together involve a possible uncertainty of one or two parts in 10000, and that the calorimetric measurements have the same accuracy. Jaeger, in quoting this work[7] gives two parts in 10000 (i.e., 8×10^{-4} joules) as the probable error. I therefore write $J_{15}'=4.1833\pm0.0008$ int. joules as the revised Jaeger and Steinwehr result. Using the results of section E, we have one int. joule $=pq^2$ abs. joule, where $pq^2=1.00041\pm0.00010$. Hence there results

$$J_{15}=(4.1833\pm0.0008)(1.00041\pm0.00010)=4.1850\pm0.0009 \text{ abs. joules.}$$

The most accurate direct determination of the *mechanical* equivalent of heat J is that resulting from the work of Laby and Hercus,[8] which has appeared since the HP was compiled. These investigators use a continuous flow calorimeter and make 23 determinations, grouped about six different mean temperatures, the temperature change in the calorimeter being always about 5°C. These are then averaged to give six determinations at different temperatures, weighted arbitrarily in rough accordance with the self consistency of the individual results within each group. Their weighting is certainly more logical in the present case than the strict least squares weighting, and is accepted by the writer. They then obtain a *weighted* average of the six results, a corresponding weighted mean temperature, and the probable error of the weighted average, by the usual least squares formula. This result is $J=4.1841\pm0.0001$ abs. joules, at 16.67°C. This method of averaging would be legitimate if the actual $f(t)$ representing J could be replaced by a straight line, over the range of temperature used, and if the residuals of all the six points, for the true $f(t)$ curve, are such as to indicate only the usual accidental errors.

A more precise method of averaging is first to adopt a curve for the variation of the specific heat of water with temperature. Such a curve is given immediately by Eq. (2). If it is desired that the specific heat at 15°C be unity, Eq. (2) is to be divided by 4.18327. Laby and Hercus, later in their article, assume that Callendar's revision of Barnes' data for the specific heat of water[9] gives the most reliable data, in terms of the thermodynamic scale. This matter will be discussed by the writer in the separate paper to which reference has been made, with the conclusion that the Jaeger and Steinwehr results, because of the later date of the work, are the more reliable. Over the limited temperature range now being discussed, the two specific heat curves differ by almost negligible amounts.

[7] W. Jaeger, HP. **9,** 493.

[8] T. H. Laby and E. O. Hercus, Phil. Trans. **A227,** 63 (1927). See also T. H. Laby, Proc. Phys. Soc. London **38,** 169 (1926) for a critical review of all previous determinations of the *electrical* equivalent of heat.

[9] See Landolt-Börnstein, 5th Ed. p. 1250, for all data on the specific heat of water. The Callendar reference is Phil. Trans. **A212,** 1 (1912).

Using Eq. (2) as a basis for a specific heat curve, I have reduced each Laby and Hercus result to the 15° calorie, the six temperatures involved being15.97, 15.88, 16.15, 16.74, 15.62, and 20.50, all in °C. It then appears that the reduced result for the 20.50°C temperature is quite inconsistent with the other five.[10] Thus the first five results yield a weighted average of 4.18526 abs. joules for the 15° cal., with a weighted mean residual of 3.0×10^{-4} joules, while the residual of the sixth determination is 22.1×10^{-4}. If the Callendar specific heat curve is used, the corresponding results are 4.18545, 3.1×10^{-4}, and 31.5×10^{-4}, respectively. An observation with a residual 7 to 10 times the mean residual should be rejected, according to any of the well known criteria of rejection. Hence I obtain from the Laby and Hercus data, $J_{15}=4.18526$ abs. joules.

Laby and Hercus discuss the accuracy of the various elements entering into the determination of their result, but make no estimate of the probable error in the final result, due to the various possible sources of error. The only error given (±0.0001) is, as noted, merely a least squares error, due to accidental variations of the individual results. Jaeger and Steinwehr, on the other hand, make no *specific* allowance for such an error, and their stated error (±0.0008) seems to be their estimate of the various possible errors. constant and otherwise, affecting the result. Both these investigations have been carried out during the past decade, with every possible experimental refinement, and it is natural to assume that the final probable error in the two cases cannot be greatly different. Of the 67 individual determinations by Jaeger and Steinwehr, six show very large residuals, and have been rejected by the writer, in deriving Eq. (2). The remaining 61 have a mean residual, in terms of Eq. (2), of 11.7×10^{-4} joules. In the case of Laby and Hercus, the four individual determinations made at the relatively high mean temperature, 20.50°C, have been rejected as a group, because of their very large mean residual. The remaining 19 determinations forming the other five groups have an average residual of 8.8×10^{-4} joules, assuming the weighted result 4.18526 abs. joules for the 15° calorie, and the specific heat curve given by Eq. (2). The purely accidental errors, as anticipated, are thus roughly the same for the two investigations, and in lieu of more definite evidence it seems best to assume that the *total* probable error in both cases is the same as that given by Jaeger and Steinwehr for their own work. Hence I take as the final Laby and Hercus result, $J_{15}=4.1853\pm0.0008$ abs. joules.

This is to be averaged with the reduced value 4.1850 ± 0.0009 abs. joules, resulting from the work of Jaeger and Steinwehr. Because of the greater accuracy claimed by Laby and Hercus, I adopt for the mean value 4.1852, rather than 4.1851. The probable error of the mean can however be taken slightly less than that of either determination. My final assumed result is

$$\text{one } 15^\circ \text{ calorie } (J_{15})=4.1852\pm0.0006 \text{ abs. joules.}$$

[10] This is most probably due to some constant error at the 20.5° temperature. All other observations, it should be noted, are close to 16°C mean temperature.

Using now the assumed value $pq^2 = 1.00041 \pm 0.00010$, we obtain

$$\text{one } 15° \text{ calorie } (J_{15}') = 4.1835 \pm 0.0007 \text{ int. joules,}$$

as a *slightly* more accurate value than the direct determination by Jaeger and Steinwehr of 4.1833 ± 0.0008 int. joules.

Finally, using Eq. (2) one obtains $J_{20}/J_{15} = 0.999058$. Hence one 20° calorie

$$J_{20} = 4.1813 \pm 0.0006 \text{ abs. joules}$$

$$J_{20}' = 4.1796 \pm 0.0007 \text{ int. joules.}$$

In conclusion it should be noted that the revised Jaeger and Steinwehr result, $J_{15}' = 4.1833 \pm 0.0008$ int. joules, when taken with the revised Laby and Hercus result, $J_{15} = 4.1853 \pm 0.0008$ abs. joules, gives directly one int. joule $= 1.00048 \pm 0.00027$ abs. joules. This is an entirely independent method for evaluating pq^2, but the probable error is about three times as large as that assumed for the more direct method discussed in section E. The good agreement of the two values, 1.00041 and 1.00048, is therefore probably more or less accidental. Attention should be directed also to an investigation by W. Groth[11] of the electrical and mechanical equivalents of heat. He found that within the limits of error of his work (one part in two thousand) the two quantities were identical. This agrees with our adopted values, the difference between J and J' being slightly less than one part in two thousand.

Section J

The Faraday (F)

The Faraday is defined as the quantity of electricity carried in electrolysis by one gram equivalent of any element. It is believed to be a general constant of nature, and its true value is accordingly independent of experimental conditions. According to modern ideas, each univalent ion carries a charge numerically equal to the electronic charge e. The Avogadro number N_0 gives the number of atoms (or molecules) in one gram equivalent. Hence one may define the Faraday (F) more precisely as equal to the product $N_0 \cdot e$. The fact that, as a matter of experimental practise, F can be most accurately evaluated from electrolysis, and N_0 is then evaluated from F and e, does not affect the validity of the definition.

The electrochemical equivalent of a substance is the mass associated with unit electric charge. Like the Faraday, its true value is independent of experimental conditions, and depends only on the adopted unit of charge. On the other hand, as noted in section E, we can measure only the amount of a substance deposited or released in an electrolytic cell, per unit current per second. This amount is affected by experimental conditions, and may or may not equal the electrochemical equivalent. The Faraday is then, by definition, the ratio of the gram equivalent of a substance to its electro-

[11] W. Groth, Ann. Physik **82**, 1156 (1927).

chemical equivalent. It seems to be almost universal practise to ignore the distinction between mass deposited per unit charge, and electrochemical equivalent. This practise has resulted in considerable confusion in the literature regarding the best value of certain electrochemical equivalents, and the resulting best value of the Faraday.

In spite of the above remarks, it is convenient to *assume*, for the moment, that the amount of silver deposited per unit charge in a silver voltameter, operating under the conditions laid down in defining the international ampere, is in fact the electrochemical equivalent of silver (E_{Ag}). *With this assumption*, the value of Faraday follows immediately from constants already adopted. The gram equivalent of silver, or of any univalent substance, is numerically equal to its atomic weight in grams (Ag). The amount of silver deposited in electrolysis by one international coulomb is, by definition,[1] 0.00111800 grams. Hence

$$F = \frac{\mathrm{Ag}}{0.00111800} \text{ int. coul.} \qquad (1)$$

Using our adopted value[2] of Ag, we have

$$F = \frac{107.880 \pm 0.001}{0.00111800} = 96494 \pm 1 \text{ int. coul.} \qquad (2)$$

With our adopted value[1] of q, there results

$$F = \frac{(107.880 \pm 0.001)(0.99995 \pm 0.0005)}{0.00111800} = 96489 \pm 5 \text{ abs. coul.} \qquad (3)$$

If one uses $q = 1$, as adopted by the HP[3], one gets $F = 96494$ int. coul. *or* abs. coul. This is the actual value of the Faraday adopted by Henning and Jaeger[3] If one uses $q = 0.99993$, as adopted by the ICT,[4] there results

$$F = \frac{(107.880)(0.99993)}{0.00111800} = 96487 \text{ abs. coul.} \qquad (4)$$

The ICT, however, adopts[5] $F = 96500 \pm 10$ abs. coul., which with its adopted value of q, leads to[6] $F = 96507$ int. coul. This last value, if substituted in Eq. (1) obviously requires Ag = 107.893, in direct contradiction to the known facts. The value $F = 96500 \pm 10$ abs. coul. is evidently taken from the work of Vinal and Bates,[7] and in order to understand the seeming discrepancy, it is necessary to examine in some detail this last quoted investigation, and to employ the distinction just mentioned between mass carried in electrolysis and mass deposited.

[1] See section E.
[2] See section F.
[3] Henning and Jaeger, HP **2** 502.
[4] ICT. **1**, 18.
[5] ICT. **1**, 17.
[6] ICT. **1**, 26.
[7] G. W. Vinal and S. J. Bates, Bur. Standards, Bull. **10**, 425 (1914). See page 447.

Vinal and Bates were investigating the iodine voltameter, and connected such a voltameter in series with a silver voltameter. They then measured the ratio of the mass of silver to the mass of iodine, deposited (or liberated) by the same current. *If* these masses truly represent the amounts of silver and of iodine required to carry a specified quantity of electricity, their ratio should, by Faraday's law, equal the ratio of the atomic weights of silver and iodine. The final measured value of this ratio (Ag/I) was 0.85017, with an apparent probable error of a few parts in 10^5. This result has been mentioned in section F. Combined with Ag = 107.880, it gives I = 126.892, while as discussed in section F, the best chemical value of the ratio leads to I = 126.932 ± 0.002. The Vinal and Bates value is thus in disagreement with the chemical value by about 32 parts per 10^5, and this discrepancy must be due to the fact that in the silver voltameter, or in the iodine voltameter, or in both, the measured amount of deposit does not represent the true mass carried by a given electric charge.

This question has been investigated in great detail, in the case of the silver voltameter. In particular, the possibility of inclusions in the silver deposit has been studied by various persons.[8] The final conclusion is that there is likely to be about 0.004 percent of inclusions in the silver deposit of a silver voltameter, operated under the specifications used in defining the int. amp. Applying this correction to the Vinal and Bates results, one gets Ag/I = 0.850136, accounting for only 3.4 parts per 10^5, out of the total discrepancy of 32 parts. It would therefore appear that, barring unsuspected sources of error in the silver voltameter, such as secondary chemical reactions, etc., the main cause of the above discrepancy lies in the iodine voltameter.

Vinal and Bates, working before the above facts as to inclusions were known, did not draw this conclusion. Their actual procedure was to use Eq. (1) to get $F = 96494$ int. coul. from the silver voltameter, and then to use I = 126.92 (the accepted value at that time), and a calculated electrochemical equivalent of iodine (E_I), to get a new value $F = 96514$ int. coul. from the iodine voltameter. Their value of E_I (0.00131502) was obtained by dividing 0.00111800 by their measured ratio 0.85017. They then take the arithmetic average of 96494 and 96514, or 96504, as the best value of the Faraday in int. coul. They state, finally, that the value of q (as we have defined it in section E) is uncertain, but that 96500 ± 10 abs. coul. can be considered the best working value of F. This assumed value in abs. coul. evidently corresponds to $q = 0.99996$, in close agreement with our accepted value.

Their adoption of 96504 int. coul. as the mean value of the Faraday necessarily implies the following conclusions. From Eq. (1), with $F = 96504$ int. coul., and Ag = 107.880, we obtain $E_{Ag} = 0.00111790$ grams. Hence one must conclude that although 0.00111800 grams of silver are deposited apparently by one int. coul., and serve as its definition, the true electrochemical equivalent is only 0.00111790 grams per int. coul. Using the Vinal

[8] E. B. Rosa, G. W. Vinal, and A. S. McDaniel, Bur. Standards, Bull. **10**, 475 (1914); G. W. Vinal and W. M. Bouvard, ibid **13**, 147 (1916).

and Bates value of I, we have $96504 = 126.92/E_I$. Hence $E_I = 0.00131518$, as compared to 0.00131502 deduced by them, from the amount of iodine liberated. We must thus conclude that in this case, the amount of the substance (iodine) liberated is *less* than that required to carry the charge. This may very well be the case, due to the loss of liberated iodine, and the matter is discussed by Vinal and Bates. There does not, however, seem at the present time to be experimental justification for dividing up the discrepancy practically equally between the silver and iodine voltameters. The only known fact appears to be the observed 0.004 percent inclusion in the silver deposits, and the proper correction for this would give 0.00111796 grams as the true E_{Ag} per int. coul., or, with our adopted value of q, 0.00111801 g as the true E_{Ag} per abs. coul. These values in turn lead to $F = 96498$ int. coul., or 96493 abs. coul.

In a later paper Vinal and Bouvard[8] do in fact discard entirely the iodine value, and by this same method deduce $F = 96494$ abs. coul., based on $q = 0.99996$. They however again recommend 96500 abs. coul. as the best rounded value. On the other hand, Henning and Jaeger[3] make no distinction between mass carried and mass deposited, writing $E_{Ag} = 0.00111800$ g per int. coul., with *no* indicated error. It seems evident from the work of Vinal and Bouvard that there *are* inclusions in the silver deposit, tending to make E_{Ag} too large by 4×10^{-8} g, and F too small by 4 coulombs. On the other hand, there may well be small parasitic chemical reactions in the silver voltameter, tending to decrease the value of E_{Ag} and hence to increase the value of F. Under the circumstances it seems to the writer best to *adopt* the values of F given in Eqs. (2) and (3), but to assign to E_{Ag} a probable error of 5×10^{-8}g, i.e., an error slightly greater than the measured effect of the inclusions. Then

$$F = \frac{107.880 \pm 0.001}{(1.11800 \pm 0.00005) \times 10^{-3}} = 96494 \pm 5 \text{ int. coul.} \qquad (5)$$

and

$$F = \frac{(107.880 \pm 0.001)(0.99995 \pm 0.00005)}{(1.11800 \pm 0.00005) \times 10^{-3}} = 96489 \pm 7 \text{ abs. coul.} \qquad (6)$$

Eq. (5) thus assumes $E_{Ag} = (1.11800 \pm 0.00005) \times 10^{-3}$ g per int. coul., while Eq. (6) assumes $E_{Ag} = (1.11805 \pm 0.00007) \times 10^{-3}$ g per abs. coul. Finally from Eq. (6),

$$F = 9648.9 \pm 0.7 \text{ abs. } em \text{ units}$$

$$Fc = (9648.9 \pm 0.7)(2.99796 \pm 0.00004) \times 10^{10}$$

$$= (2.89270 \pm 0.00021) \times 10^{14} \text{ abs. } es \text{ units.}$$

Section K

The Electronic Charge (e)

A relatively large number of important constants have values which depend directly on the value assigned to the electronic charge e, and in most cases the final probable error in these values is due mainly to the error in e.

From this standpoint the electronic charge may be considered the most important of the general physical constants, and it is therefore highly desirable that its value should be determined in many different ways, and by many different persons. Until very recently the situation has been the reverse of this. Only one really precision method for the evaluation of e was known, and the precision work, by this method, had been carried out by a single individual. Under such circumstances it is very fortunate that the investigations referred to represent a masterpiece of experimental technique, in which every possible cause of error, systematic or accidental, was studied in minute detail, and after many years of work reduced to a final minimum. Millikan's investigations on this subject, by the oil-drop method, extend over more than a decade, the latest value of e having been published[1] in 1917. Because of the great importance of this resulting value of e, and because of higher values which have recently been obtained, the writer has investigated the matter in more than usual detail.

Millikan found that if the viscosity of air is taken as constant, in Stokes' law of fall, the apparent value of e is a function of the radius of the drop and of the pressure of the air. The true value of e can be found by assuming a modification of Stokes' law. If the pressure p or the radius of the drop a is sufficiently large, the final equation for determining the true e is

$$e_1^{2/3} = e^{2/3}(1 + b/pa) \tag{1}$$

where e_1 is the apparent value, e the true value, and b an undetermined constant. This equation can be written

$$y = a_0 + a_1 x \tag{2}$$

when

$$x = 1/pa, \quad y = e_1^{2/3}, \quad a_0 = e^{2/3}, \quad a_1 = e^{2/3} \cdot b.$$

Hence if $e_1^{2/3}$ is plotted against $1/pa$, the resulting graph should be linear, and its intercept on the y axis gives $e^{2/3}$, leading to the desired quantity.

Millikan found that for values of $1/pa$ less than about 700 (p in cm Hg, a in cm), the resulting graph was definitely linear. For larger values of $1/pa$, the slope of the graph increases to a new larger value, and then remains constant. The complete law for the fall of such drops in air was discussed[2] in 1923. Only that part of the curve corresponding to $1/pa$ less than 700 was used in precise dete minations of e. The 1917 value of e was deduced from data on 25 oil-drops, each drop giving one point on the graph. The 25 observations form a beautifully consistent set of data,[3] lying on a linear graph. The least squares solution, as calculated by the writer, gives for the intercept $a_0 = (61.111 \pm 0.032) \times 10^{-8}$. The plotted data are however

[1] R. A. Millikan, Phil. Mag. (6) **34,** 1 (1917). For previous work by Millikan see Phys. Rev. **29,** 60 (1909), Phil. Mag. **19,** 209 (1910), Phys. Rev. **32,** 349 (1911), Phys. Rev. **2**; 109 (1913), and Physik. Zeits. **14,** 796 (1913). See also R. A. Millikan "The Electron," Chap. V.

[2] R. A. Millikan, Phys. Rev. **22,** 1 (1923).

[3] See R. A. Millikan, Phil. Mag (6) **34,** 12 (1917), or "The Electron" page 121 (first edition). for the plot of these data.

based on the 1913 value (0.0001824) for the viscosity of air. The value of $a_0(=e^{2/3})$ is directly proportional to the assumed viscosity. With the improved 1917 value of the viscosity (0.00018227), a_0 becomes 61.0676×10^{-8}. Hence $e=a_0^{3/2}=(4.7721\pm0.0038)\times10^{-10}$ *es* units.

Millikan stars 18 of the points, these having been taken when the conditions of observation were considered as perfect as possible. These 18 drops give $a_0=61.121\pm0.038$ (or $e=4.7733\pm0.0045$, based on the 1917 viscosity). As a matter of fact, these 18 drops deviate from the best straight line more than do the other 7. Thus the standard deviation (mean error of a single observation) of the 25 drops is 0.121×10^{-8}, while for the 18 drops it is 0.123×10^{-8}, each referred to its own least squares solution. The drops of smaller radius fall more slowly, and therefore can be more accurately timed. This might make them appear more reliable. Actually they are less reliable, due presumably to other errors. Thus 13 smaller drops whose radii lie between 12.17 and 15.92×10^{-5} cm, have a standard deviation of 0.134, when considered as part of the 25 drops. This is definitely larger than the 0.121 average of the entire 25. A least squares solution, based on these 13 drops alone gives $a_0=61.143\pm0.050$, and a standard deviation of 0.132. This standard deviation is so close to 0.134 that we can conclude that the 13 drops fit on the graph of the entire 25 essentially as well as on a graph designed to fit them alone. On the other hand, the 12 larger drops (for which $a=16.57$ to 23.40×10^{-5} cm) give for the least squares solution, $a_0=61.078\pm0.045$, and a standard deviation 0.117. They are thus definitely more reliable than the smaller drops, as judged by their mutual consistency. The resulting value of e, reduced to the 1917 viscosity, is 4.7759 ± 0.0058 for the 13 smaller drops, and 4.7683 ±0.0053 for the larger drops. The weighted mean is 4.7718, in essential agreement with the value (4.7721) obtained from all 25 drops. This of course is just what we should expect.

The average deviation from the average of the two results (for small and large drops respectively) is 0.0038, and this is much less than the probable error of either. This is an analytic proof of the contention made by Millikan in 1913 and again in 1917, that the true value of e (intercept of the graph) is not in any way a function of the radius of the drop. The above results also indicate that the larger drops are, if anything, more reliable than the smaller. If the larger are given a higher weight, the resulting value of e would lie between 4.772 and 4.768, according to the weighting. It has been noted already that the 18 starred drops are slightly less consistent than the others. The final conclusion is then that there is no particular reason for giving different weights to the different drops, and that any such weighting, if made, would slightly lower e. I therefore take 4.772×10^{-10} *es* units as the best result of the 1917 work.

In 1913 data on 58 drops were obtained. These show practically the same consistency as the 25 drops just considered, the standard deviation being 0.135. Millikan actually used, in evaluating e, the 23 drops (out of 58) of smallest $1/pa$. These 23 are more consistent, having a standard

deviation of only 0.092. They lead to $e = 4.7665 \pm 0.0058$, based on the 1917 viscosity, while the entire 58 drops give 4.7703 ± 0.0022. This last figure might appear to be a more reliable value of e than that obtained in 1917, but such a conclusion ignores the errors due to other factors. In 1913 Millikan estimated that there were four such factors, each with a maximum uncertainty of 0.1 percent. In 1917 he estimated that there remained only two such factors of any consequence, each with a maximum of uncertainty of only 0.05 percent. His final estimate, in 1917, for the maximum of uncertainty in e is 0.1 percent, based mainly on these two factors. The above calculations show, however, that there is a *probable* error of 0.08 percent (± 0.0038) in the 1917 value of e, due merely to accidental errors. The final maximum uncertainty is therefore several times as large. I estimate that the final *probable error* is about 0.1 percent, and write $e = (4.772 \pm 0.005) \times 10^{-10}$ *es* units.

This value is now subject to two further corrections. The equation for e (or e_1) can be put in the form

$$e = k/F \tag{3}$$

where F is the electric field intensity, and k represents a number of factors, all mechanical in nature. F is measured directly in int. volts per cm, using a calibrated Weston cell. In reducing the result to *es* units per cm, Millikan used $c = 2.999 \times 10^{10}$ cm·sec^{-1}, and made no distinction between international and absolute electrical units. Since the time of his work it has been shown definitely that the int. volt differs from the abs. volt by an appreciable amount, as discussed in section E. We have also now the new value, $c = 2.99796$. Both the resulting corrections tend to lower e. That due to the change in c is of course obvious. It lowers e in just the proportion 2.99796/2.999, and so changes e from 4.772 to 4.770. The other change, due to the electrical units, seems to have been overlooked by everyone. Thus Henning and Jaeger, in the HP[4], state explicitly that Millikan used international volts, and that the resulting e is therefore in int. coulombs. The difference between int. coul. and abs. coul. being negligible, they therefore make no correction. This conclusion is however incorrect. Because the electrical potential forces the charged drops against the viscosity of air, instead of against *electrical* resistance, one has *only* electric voltage coming into the calculations, as shown by Eq. (3). It has been shown in section E that one int. volt $= 1.00046 \pm 0.00005$ abs. volts. Hence the true value of F, in abs. volts, is larger and the true value of e, in *abs. es* units is smaller, by just this ratio. Hence, finally, the value of e is reduced[5] from 4.770 to 4.768. Since the error in each of these corrections is negligible, the final result is

[4] HP **2,** 503.

[5] These two corrections to e, due to the change in c, and to the electrical units, were first noted by the writer in Sept. 1928, and transmitted immediately to Professor Millikan who agreed with them. They were later presented to the Am. Phys. Soc., at the Pasadena meeting in December 1928 (Phys. Rev. **33,** 265, 1929).

$e = (4.768 \pm 0.005) \times 10^{-10}$ *abs. es* units. This, the writer believes, is the most reliable value that can be deduced from Millikan's oil-drop work.

Very recently there has been devised an entirely different method for measuring e, which bids fair to become more accurate than the oil-drop method. The two results by this method which have already been published are, however, apparently less reliable than the oil-drop value. This new method measures directly the Avogadro number N_0, and from this and the known value of the Faraday, e immediately follows.[6] The method utilizes the absolute wave-lengths of x-ray lines, determined with an ordinary ruled grating at grazing incidence, as compared with the wave-lengths determined with a crystal grating. The latter values are given by

$$\lambda = 2d \cdot \sin\theta \tag{4}$$

where d is the grating space. It has been pointed out by Siegbahn,[7] and by Compton, Beets and DeFoe,[8] that in order to obtain the *true* x-ray wave-length λ, it is necessary to use an *effective* grating space d in Eq. (4). Such an effective grating space automatically corrects for the actual refraction of the x-rays at the crystal surface. For first order spectra, and for the high frequencies of ordinary x-rays, the *true* grating space d' is connected with the effective space d by the relation

$$d = d'(1 - 0.000135). \tag{5}$$

In evaluating λ, Siegbahn uses for the calcite grating space $d = 3.02904 \times 10^{-8}$ cm, at 18°C. This is a more or less arbitrary value, based on the assumption that d for rock-salt, at 18°C, is 2.81400×10^{-8} cm. We shall denote by d''_{18} this 3.02904 value, and by λ'' the resulting wave-length. Hence

$$\lambda'' = 2d_{18}'' \sin\theta. \tag{6}$$

From Eqs. (4) and (6)

$$\lambda/\lambda'' = d_{18}/d_{18}'' \tag{7}$$

in which λ is the *true* wave-length of an x-ray, as determined with a ruled grating, d_{18} the effective grating space of calcite at 18°C, λ'' the supposed true wave-length resulting from measurements with a calcite crystal with d''_{18} as an assumed effective grating space at 18°C. Thus knowing λ'', d''_{18}, and λ, one calculates d_{18}. Then from Eq. (5) we obtain d'_{18}, the true grating space of calcite. The temperature coefficient of calcite[9] is 1.04×10^{-5} and the value of d'_{20} is accordingly 2.08×10^{-5} larger. This 20° value of d' is given theoretically by the following formula,

$$d_{20}' = \left\{ \frac{nM}{\rho N_0 \phi(\beta)} \right\}^{1/3} \tag{8}$$

in which $n = 1/2$, M = molecular weight of calcite ($CaCO_3$), ρ = density of calcite at 20° C, $\phi(\beta)$ = a geometrical constant depending on the crystal

[6] See section J.
[7] M. Siegbahn, "Spectroscopy of X-rays" p. 26.
[8] A. H. Compton, H. N. Beets, and O. K. DeFoe, Phys. Rev. **25**, 625 (1925).
[9] M. Siegbahn, reference 7, p. 85.

structure of calcite, N_0 = Avogadro's number $= F/e$ where F = Faraday. Thus knowing d'_{20} we can calculate N_0 and finally e.

The best value of M follows directly from the atomic weights adopted in section F. It is

$$M = 3\times 16.000 + (12.003 \pm 0.001) + (40.075 \pm 0.005) = 100.078 \pm 0.005 \quad (9)$$

The best value of ρ is 2.7102 ± 0.0004 g·cm^{-3}, at 20°C, as determined by DeFoe and Compton.[10] The best value of $\phi(\beta)$ is 1.09630 ± 0.00007 at 20°C, as determined by Beets.[11] The adopted value of the Farady[12] is $(2.89270 \pm 0.00020) \times 10^{14}$ abs. *es* units. Solving Eq. (8) for e, and making the above substitutions, we have

$$e = \frac{\rho F \phi(\beta)(d')^3}{nM} = (1.7176 \pm 0.0003) \times 10^{13} \cdot (d_{20}')^3. \quad (10)$$

The two published determinations of d_{18}, based on absolute x-ray wave-lengths, are by Bäcklin,[13] and Wadlund.[14] Using Eq. (7), Wadlund obtains 1.5373 ± 0.0008A for the $K\alpha_1$ line of Cu, and this, combined with Siegbahn's values of d''_{18} and λ'', gives $d_{18} = (3.0290 \pm 0.0016) \times 10^{-8}$ cm. The corresponding value of d_{20} is then 3.02906, and the true grating space d'_{20} is by Eq. (5), $(3.0295 \pm 0.0016) \times 10^{-8}$ cm. This value[15] is to be substituted in Eq. (10). It gives $e = (1.7176 \pm 0.0003) \times 10^{13} (3.0295 \pm 0.0016)^3 \times 10^{-24} = (4.7757 \pm 0.0076) \times 10^{-10}$ abs. *es* units. Wadlund, using other values of the auxiliary constants, obtained 4.774 ± 0.007. This is not as accurate as the oil-drop value, and Professor A. H. Compton writes me that later work indicates that Wadlund's published probable error may be a slight underestimate.

It is very difficult to appraise the work of Bäcklin,[13] especially as regards its accuracy. He gets 8.333 ± 0.008A for the absolute wave-length of the Al $K\alpha$ line. Comparing this with an unpublished result by A.Larsson (8.3229 ± 0.0008A), obtained with a crystal, Bäcklin obtains $d_{18} = 3.033 \pm 0.003$A. This gives $d'_{20} = 3.03347$A, and by Eq. (10), $e = (4.794 \pm 0.015) \times 10^{-10}$ abs. *es* units. This value is 0.55 percent higher than the oil-drop result. His own result, with different values of the auxiliary constants, is 4.793 ± 0.015.

Before discussing further the absolute value, a word is necessary as to the errors. Throughout his paper, Bäcklin calculates the *mean error* of a *single* observation (i.e., the standard deviation) and seems to consider

[10] O. K. DeFoe and A. H. Compton, Phys. Rev. **25,** 618 (1925).

[11] H. N. Beets, Phys. Rev. **25,** 621 (1925).

[12] See section J.

[13] E. Bäcklin "Absolute Wellenlängenbestimmungen der Röntgenstrahlen," Uppsala Dissertation, 1928.

[14] A. P. R. Wadlund, Proc. Nat. Acad. Sci. **14,** 588 (1928), and Phys. Rev. **32,** 841 (1928).

[15] Both Wadlund and Bäcklin omit these corrections and thus use their observed d_{18} as the value of d'_{20}.

that this is the *limit of error* of the resulting *average* value. For the 15 to 25 observations he is normally handling, such a standard deviation is several times the probable error of the average. It therefore *does* represent more nearly a reasonable *limit* of error than a *probable* error. On the other hand, in getting the final error in λ, he takes the square root of the sum of the squares of the individual limits of error. As stated in the Introduction one should in this case take the straight sum, to get the limit of error of the result. This sum is 0.17 percent for the error in λ, instead of his 0.10 percent, and so 0.51 percent in e. This gives $e=4.794\pm0.024$, where the *limit of error* is indicated. The probable error is then not over ±0.006, so that the apparent accuracy is practically that of Millikan's work. In fact Bäcklin compares the average residual of his observations with that of Millikan, and finds his definitely smaller, and also finds his residuals better distributed on the Gaussian error curve. In terms of the resulting value of e, his average residual is 0.12 percent, with 29 observations. The residuals for the oil-drop method, used by Bäcklin, are not based on a least squares solution of the data, but the average residual which he gives, 0.25 percent, is essentially the same as that given by the correct least squares solution of the 25 drops used by Millikan in 1917. His final error is however slightly larger than that for the oil-drop work, due to the fact that his other sources of error are larger. He rightly concludes that as far as internal consistency is concerned, the x-ray method is more accurate, and the problem thus becomes that of eliminating the various sources of systematic error.

In the preliminary work on this paper, I calculated each general constant quite independent of the relations between the various constants. The values so obtained were found to be unexpectedly satisfactory, when certain known or proposed relations were tested. In the subsequent extensive revisions, these relations have been kept in mind, and a deliberate attempt has been made to retain a *system* of constants which should be as satisfactory as possible. The most important of the relations between general constants is that given by the Rydberg constant, which according to Bohr is a simple function of e, e/m, h, and c. This relation is discussed in section O, in connection with the finally adopted values of these four constants. In anticipation of this discussion it may be said that any value of e as high as that of Bäcklin could hardly be reconciled with the observed values of h and e/m.

The investigation by Bäcklin constitutes a pioneer piece of work, and it is quite likely, as such, to contain various unsuspected sources of systematic error. If the three values of e (4.768 from Millikan's oil-drop work, 4.776 by Wadlund, and 4.794 by Bäcklin) are weighted according to the apparent probable error of each, the weighted average will still be suspiciously high. The very thorough examination made by the writer of the actual value of e and its probable error, as deduced from oil-drop work, was carried out mainly because of the inconsistency between that value and the result published by Bäcklin. In view of all the above circumstances, the writer has finally decided to reject the Bäcklin value, and to use the weighted

mean of the remaining two values, viz. 4.768 ± 0.005 and 4.776 ± 0.008. This mean is 4.770, and as usual I adopt as its probable error the smaller of the two individual errors, rather than that given by least squares, since the latter is rather meaningless when only two observations are concerned. The finally adopted value is then

$$e = (4.770 \pm 0.005) \times 10^{-10} \text{ abs. es units.}$$

In conclusion attention should be called to the fact that a more accurate determination of e, from x-ray measurements, is now under way, and it is to be hoped that this work will clear up the present somewhat uncertain situation. This is discussed further in Section O.

Section L

The Specific Charge of the Electron (e/m)

A very complete and critical account of all work on the measurement of e/m, up to 1919, has been given by Bestelmeyer.[1] His final conclusion is that $e/m = (1.76 \pm 0.02) \times 10^7$ *em* units. A more recent discussion of all experimental results is that by Gerlach,[2] who concludes that $e/m = 1.766 \times 10^7$ *em* units. The question is discussed very briefly by Henning and Jaeger,[3] who however adopt Gerlach's value. The ICT[4] adopts 1.769 ± 0.003.

There is an enormous amount of experimental material on this subject. The latest work, however, so greatly exceeds in accuracy all the preceding that it seems legitimate to confine the discussion to these new results. The value of e/m has been obtained with considerable accuracy by three distinct methods, (*a*) deflection of electrons in electric and magnetic fields, (*b*) Zeeman effect, (*c*) fine structure and relative wave-lengths of H and He^+ spectral lines. It may be obtained also from Bohr's theoretical expression for the Rydberg constant, R_∞, provided one assumes the value of e and of h. This last method is, however, not as accurate as the preceding. A fifth, and very interesting method for obtaining e/m involves the Compton shift. This also is as yet a relatively inaccurate method. The last two methods are discussed in sections O and M respectively.

The latest and most accurate work with method (*a*) is that by Wolf.[5] In this work a new method, first suggested and employed by Busch,[6] is carried out with every possible experimental refinement. The essential point in the method is the employment of a longitudinal instead of a transverse magnetic field. The electron velocity is calculated from the potential fall. Wolf makes an exhaustive study of the possible errors, and finally

[1] Marx, Handbuch d. Radiologie V, pp. 1–82 (1919).
[2] HP. **22**, 41–82 (1926).
[3] HP. **2**, 504–506.
[4] ICT. **1**, 17.
[5] F. Wolf, Ann. Physik **83**, 849 (1927).
[6] H. Busch, Physik. Zeits. **23**, 438 (1922).

concludes that $e/m = (1.7679 \pm 0.0018) \times 10^7$ *em* units. So far as the writer can judge, the assumed probable error is entirely justified.

The actual magnitude, 1.7679, should however be corrected for the difference between the int. and abs. electrical units (section E). The equation used by Wolf is

$$\frac{e}{m} = \frac{8\pi^2 V}{H^2 l^2} \cos^2 \alpha \tag{1}$$

where V is the potential drop and H the magnetic field. Wolf does not state whether he corrected from int. to abs. units, but presumably, in conformity with current practise, he did not. If V' and H' denote the measured quantities in int. units, then

$$\frac{e}{m} = \frac{8\pi^2 (V' pq)}{(H'q)^2 l^2} \cos^2 \alpha \tag{2}$$

Hence his value of e/m should be multiplied by p/q and this, by section E, is 1.00056. Therefore his true result is $(1.7689 \pm 0.0018) \times 10^7$ abs. *em* units.

In the determination of e/m by means of the Zeeman effect, the basic equation is

$$\frac{e}{m} = \frac{4\pi c \Delta\lambda}{n H' q \lambda^2} \tag{3}$$

where e/m is in *em* units,

c = velocity of light in vacuum, in cm $\cdot$ sec^{-1},

n = index of refraction of the medium (air) in which $\Delta\lambda$ and λ are measured,

H' = intensity of the magnetic field in int. gauss,

$H'q$ = intensity of the magnetic field in abs. gauss,

$\Delta\lambda$ = separation of either component of a normal triplet of wavelength λ from the undisturbed position of the line, $\Delta\lambda$ and λ being measured in cm.

By section E, $q = 0.99995 \pm 0.00005$, and can be taken as unity without appreciable error.

The most recent and accurate work, using this method, is that by Babcock.[7] In his investigation a large number of spectral lines (116 in all) were employed. Nearly all of them showed a complex Zeeman effect pattern. For determining e/m it was necessary to assume the Runge denominator of each line. In cases where this is small, it was known with certainty. In some cases it was large and rather uncertain. Babcock states,[8] "It will be observed, however, that these large denominators play a minor part in the present investigation." In spite of this remark, his work has been

[7] H. D. Babcock, Astrophys. J. **58**, 149 (1923).

[8] H. D. Babcock, reference 7, p. 155.

criticised,[9] and Gerlach, in his final table[10] of values of e/m, omits Babcock's result. It appears to the writer that the criticism is unjustified, and at his suggestion, Babcock has recently recalculated his data, omitting all Zeeman patterns which are in any way doubtful.

His previously published result, assuming $c=2.9986\times10^{10}$, is $e/m=(1.761\pm0.001)\times10^{7}$ *em* units, in which the error quoted is merely the observational error, as determined by least squares. Babcock estimates his probable uncertainty, when all sources of error are considered, as not more than two or three parts in 1800. This it seems to the writer is a very conservative estimate. The new result,[11] based on 48 lines for which the Zeeman pattern is definitely established, and using $c=2.99796\times10^{10}$, is 1.7606 ± 0.0012, where again the error is purely observational. The difference between the two values is just that produced by the change in the value of c. Hence the omission of the lines of doubtful Zeeman pattern makes no change in the result. Babcock considers that his latest value is correct to about one part in a thousand. I therefore write $e/m=(1.761\pm0.002)\times10^{7}$ abs. *em* units as the best result from Zeeman effect.

The latest and most accurate work on e/m, using method (c), is that by Houston.[12] This method is based on the Bohr-Sommerfeld model of an atom consisting of a positive nucleus and *one* encircling electron (moving in elliptic or circular orbits). As examples of such an atom we have hydrogen (H) and singly ionized helium (He^{+}). In order to determine e/m, we must evaluate the so-called Rydberg constant for hydrogen (R_H) and for ionized helium (R_{He}). As will appear presently, practically the entire error in e/m is merely the error in the *difference* $R_{He}-R_H$.

The pioneer work on e/m, by this method, is that performed by Paschen.[13] He obtained $R_H=109677.69\pm0.06$ cm^{-1}, $R_{He}=109722.14\pm0.04$ cm^{-1}. This gives $e/m=1.768\pm0.003$, using his own values and assumed errors for R_H and R_{He}, but the present accepted values and errors for H, IIe, and F. The writer,[14] using better data and more recent theoretical developments, later calculated $R_H=109677.7\pm0.2$. The argument in this last quoted paper shows that Paschen's estimate of error in the case of R_H is much too small, and the entire question, in its relation to the value of e/m was later discussed.[15] In this last article it was concluded that the method here considered gives $e/m=1.758\pm0.009$. The recent investigation by Houston, which assumes the latest theoretical developments, is so much more accurate than the work just mentioned that it alone will be considered in detail.

[9] Back-Lande "Zeemaneffekt" p. 140 (1925).

[10] HP. **22,** 81.

[11] H. D. Babcock, Phys Rev. **33,** 268A (1929) and Astrophys. J. **69,** 43 (1929).

[12] W. V. Houston, Phys. Rev. **30,** 608 (1927). For his earlier work see Astrophys. J. **64,** 81 (1926), and Phys. Rev. **29,** 749 (1927) (abstract 7).

[13] F. Paschen, Ann. Physik (4) **50,** 901 (1916).

[14] R. T. Birge, Phys. Rev. **17,** 589 (1921).

[15] R. T. Birge, Nature **111,** 287 (1923).

From Bohr's theory we have

$$R_H = \frac{R_\infty}{1+m/m_H} \qquad (4) \qquad R_{He} = \frac{R_\infty}{1+m/m_{He}} \qquad (5)$$

where R_∞ = Rydberg constant corresponding to a nucleus of infinite mass,
m = mass of electron,
m_H = mass of *nucleus* of hydrogen atom
m_{He} = mass of *nucleus* of helium atom.

For convenience, m, m_H, and m_{He} are here measured in atomic weight units. From Eqs. (4) and (5) we obtain

$$\frac{m}{m_H} = \frac{(R_{He}-R_H)(m_{He}+m)}{R_H(m_{He}-m_H)} \cdot \qquad (6)$$

Houston[12] writes and uses $(m_{He}+m)$ incorrectly as $(m_{He}-m)$. In order to use directly the atomic weights, H and He, we write $m_H = H-m$, $m_{He} = He-2m$, thus obtaining

$$m = \frac{(R_{He}-R_H)(He-m)(H-m)}{R_H(He-H-m)} \cdot \qquad (7)$$

The value of m, as used on the right side of the equation, need not be known with great accuracy. The equation then gives an *accurate* value of m, where now m *must* be expressed numerically as the "atomic weight" of an electron.

From the definition of the Faraday we have[16]

$$\frac{e}{m} = \frac{F}{m} = \frac{FR_H(He-H-m)}{(R_{He}-R_H)(He-m)(H-m)} \qquad (8)$$

Houston's new experimental results are

$$R_{He} = 109722.403 \pm 0.004 \text{ cm}^{-1}, \qquad R_H = 109677.759 \pm 0.008 \text{ cm}^{-1} \cdot$$

The stated errors are purely least squares probable errors. Houston remarks that he believes the *relative* values of R_{He} and R_H are correct to 0.02, although the *absolute* error in each may be about 0.05,[17] and this seems to the writer a reasonable estimate of error.

The actual value of R_H is obtained from the strong member of the long wave-length component of $H\alpha$, assuming in accordance with the new wave-mechanics that this component consists of one strong and two weak

[16] In Eq. (8), $e/m = F/m$, the two m's are not numerically the same. e/m is merely a ratio, but if e = electronic charge, m = electronic mass in grams. In F/m, m is necessarily the atomic weight of an electron, as used in Eq. (7).

[17] The absolute values of R_H and R_{He}, depend on the absolute wave-lengths of the various spectral lines involved. These in turn depend on the wave-length of the red Cd line measured in cm. The wave-length of this line at 15°C, A_n, is 6438.4696 I.A., *defining* the International Angstrom unit (I.A.). An I.A. unit is intended to be 10^{-8} cm, and the probable error is of the order of one part in several million.

members.[18] Houston's microphotometric curve of the long wave-length component shows no asymmetry. The calculated correction to be made to the observed center of gravity in order to obtain the strong member is only 0.0056 cm^{-1}. This amounts to 0.04 cm^{-1} in R_H. In a similar way the correction in the case of $H\beta$ corresponds to 0.013 cm^{-1} in R_H. For R_{He} there is no such correction, since the desired lines are measured directly. Houston's values of R_H, resulting from $H\alpha$ and $H\beta$, differ by only 0.021 cm^{-1}. This tends to confirm the applied corrections, and it seems to the writer that these corrections must be essentially right, and that the resulting $(R_{He}-R_H)$ can scarcely contain unknown errors of more than 0.01 or 0.02 cm^{-1} I therefore assume

$$R_{He}-R_H=44.644\pm0.020 \tag{9}$$

where the second doubtful figure is carried mainly in order to get a more exact comparison of the various possible results.

Houston used $m=5.4\times10^{-4}$, $He=4.0001$, $H=1.0077$, $F=96470$ abs. coulombs, and obtained $e/m=(1.7606\pm0.0010)\times10^7$ *em* units. Using his constants and the correct formula (Eq. 8) the result is 1.7603. The error in his formula is therefore almost negligible. The values of the necessary auxiliary constants adopted in this paper are

$H=1.00777\pm0.00002$ $\qquad$ $He=4.0022\pm0.0004$

$F=96489\pm7$ abs. coul. $=9648.9\pm0.7$ abs. *em* units.

m is given by Eq. (8) as

$$m=\frac{F}{e/m}=\frac{9648.9}{1.761\times10^7}=5.48\times10^{-4}. \tag{10}$$

The entire probable error in e/m, due to errors in all factors of Eq. (8), aside from $(R_{He}-R_H)$, is less than 0.01 percent and so is entirely negligible compared to the error in $(R_{He}-R_H)$.[19]

Using Houston's value of R_H, and of $R_{He}-R_H$ (Eq. 9), together with the values of H, He, etc., just quoted, we obtain from Eq. (8) $e/m=(1.7608\pm0.0008)\times10^7$ abs. *em* units. This value of e/m thus agrees with that obtained by Babcock. Summarizing the results we find

$e/m=1.769\pm0.002$ from deflection experiments,

$=1.761\pm0.002$ from Zeeman effect,

$=1.761\pm0.001$ from H and He spectra.

The discrepancy between the first result and the last two is four times the probable error of the first, and according to the Gaussian error curve, there is only one chance in 143 of this occurring. Hence the discrepancy seems to be real.

[18] See W. V. Houston, reference 12.

[19] It may be noted that in Eq. (8), He enters in such a manner that the combined proportional error in $(He-H-m)/(He-m)$ is only one-third the proportional error in He itself.

Now the last two results constitute measurements of e/m for electrons *inside of an atom*, based upon the quantum theory of atomic structure. The first result is the measurement of e/m for electrons *in free space*. The figures thus point to the startling conclusion that the e/m of an electron is less when it is *inside* an atom than when it is *outside*. If this conclusion seems unacceptable, then it would appear that there is some general error in the equations of the quantum theory of atomic structure. The final alternative is that there is some unknown general error in all the deflection experiments. No matter what may be the cause of the discrepancy, the very fact of its existence appears to the writer to be of profound significance. Under the circumstances, it seems to be necessary to assume two different values of e/m, one to be used in all cases involving atomic structure, and the other involving free electrons. These may be called the spectroscopic and the deflection values of e/m. Hence

e/m (spectroscopic) $= (1.761 \pm 0.001) \times 10^7$ abs. *em* units per gram,

e/m (deflection) $= (1.769 \pm 0.002) \times 10^7$ abs. *em* units per gram.

Using the adopted value of c,

$(e/m)c$ (spectroscopic) $= (5.279 \pm 0.003) \times 10^{17}$ abs. *es* units per gram,

$(e/m)c$ (deflection) $= (5.303 \pm 0.006) \times 10^{17}$ abs. *es* units per gram.

Section M

The Planck Constant h.

The Planck constant h has been evaluated in a number of different ways. There is however some difference of opinion as to the relative accuracy of the different results, and some of the results are more or less incompatible. Hence a satisfactory evaluation of this fundamental constant is rather difficult.

The first attempt to obtain a most probable value of h, from the results of all seven methods by which it may be evaluated, was made by the writer[1] in 1919. The value found was $(6.5543 \pm 0.0025) \times 10^{-27}$ erg · sec, the stated error being merely the least squares probable error, deduced by the usual formula from the seven separate values, each weighted according to its probable error. This error has been criticized by Ladenburg[2] as far too small. It is however *not* the final error since, as clearly stated in my article, one must add to it an error somewhat greater than the proportional error in e. This latter quantity occurs with some positive power (unity to two) in every known method for obtaining h. This makes the total probable error more nearly ± 0.01. The writer's 1919 evaluation of h has been adopted by the ICT, but the probable error is accidentally given as ± 0.001, instead of ± 0.01.

In 1920 Ladenburg[2] wrote a detailed article on the evaluation of h, in which several of the writer's conclusions were criticized. His own result in that article was 6.54 ± 0.01. In 1925 Ladenburg wrote another article on this subject, for the HP.[3] He then concludes that $h = 6.547$, which value he rounds off

[1] R. T. Birge, Phys. Rev. **14,** 361 (1919).

[2] R. Ladenburg, Jahr. Radioakt. und Elektronik **17,** 93 (1920).

[3] HP **23,** 279–306.

to 6.55 ± 0.01. Henning and Jaeger[4] also briefly discuss the most probable value of h, and adopt 6.55, although no reference is made to Ladenburg's article in another volume. Both of Ladenburg's articles give rather full details of the theory and the experimental facts concerning the various methods for determining h.

Let us now consider briefly the present evidence on this subject. In obtaining the average value I shall exclude the Lewis and Adam's theory of ultimate rational units, and will discuss that separately in section O. This leaves six methods for evaluating h.

(a) *Bohr's Formula for the Rydberg Constant.* Bohr's theory of the hydrogen atom leads to the equation

$$R_\infty = \frac{2\pi^2 e^5}{h^3 c^2 e/m} \tag{1}$$

in which R_∞ is the Rydberg constant for infinite mass, expressed as usual in cm^{-1} units, while e, the electronic charge, is as usual in abs. *es* units, and e/m in abs. *em* units. R_∞ is derived[5] from the observed R_H by the equation

$$R_\infty = R_H\left(1+\frac{m}{m_H}\right) = R_H\left(1+\frac{F}{(e/m)(H-m)}\right). \tag{2}$$

From section L we have

$$R_\infty = (109677.759)\left(1+\frac{9648.9}{1.761\times 10^7(1.00722)}\right)$$

$$= 109677.759+59.665 = 109737.424\ \text{cm}^{-1}. \tag{3}$$

The probable error[6] in R_∞ is about 0.06 cm^{-1}. In absolute units we have $R_\infty \cdot c = (3.28988 \pm 0.00004)\times 10^{15}\text{sec}^{-1}$. Substituting in Eq. (1) the spectroscopic value of e/m, since we are here dealing with spectroscopic data, and the values found for e, c, and R_∞, we obtain

$$h = \left[\frac{2\pi^2(4.770\times 10^{-10})^5}{109737.42(2.99796\times 10^{10})^2(1.761\times 10^7)}\right]^{1/3}$$

$$= 6.54713\times 10^{-27}\ \text{erg}\cdot\text{sec}. \tag{4}$$

With the adopted probable errors of R_∞, e, c, and e/m, the resulting probable error in h is 0.011. It should be noticed that the errors in c, R_∞ and e/m, as affecting h, are negligible in comparison with the error in e. The result of this method is then

$$h = (6.547 \pm 0.011)\times 10^{-27}\ \text{erg}\cdot\text{sec}. \tag{5}$$

[4] HP **2**, 510–511.

[5] Using Eqs. (4) and (8) of section L.

[6] From section L the probable error in the absolute value of R_H is 0.05 cm^{-1} and that in the second term of Eq. (3) is 0.03 cm^{-1}.

It may be noted, in passing, that after adopting a weighted mean value of h, Eq. (1) becomes a method for calculating, indirectly, the value of R_∞. On the other hand, using the directly determined value of R_∞, Eq. (1) becomes a means for calculating e/m. Both these calculations are given in Table b of section N.

(b) *Ionization Potentials.* In 1919 the author had available thirteen values of ionization and resonance potentials, each of which could be used to calculate h. Since then many more such potentials have been obtained. The probable error in each is rather large, because of a lack of rigidly controlled conditions. We now, however, have one really accurate determination of an ionization potential, obtained with electrons of carefully controlled velocity. This is Lawrence' determination[7] of the ionization potential of Hg. His final conclusion is that this value equals 10.40 ± 0.02 int. volts. The value 10.399 ± 0.007 used by the writer[8] in a brief discussion of h and other constants, was a preliminary result,[9] in which the stated error was merely that deduced from internal consistency of the data.

The equation for obtaining h is

$$h\nu = eV \tag{6}$$

in which all quantities are in absolute units. The observed potential (V') is is always in int. volts. The potential in abs. *es* units is then

$$V = pq\ V'\ 10^8/c \tag{7}$$

p and q having been evaluated in section E. The spectral frequency ν (in sec^{-1}) is obtained always from the wave-length λ, in cm. Hence $\nu'(\text{cm}^{-1}) = 1/\lambda$, and $\nu = c/\lambda$. Eq. (6) may then be written

$$\frac{h}{e} = \frac{pqV'10^8}{c^2\nu'} = \frac{pqV'\lambda 10^8}{c^2}\,. \tag{8}$$

It seems to be quite customary to assume that $V_{es} = V'_{\text{volts}}/300$, and to write Eq. (8) as[10]

$$h/e = V'\lambda/300c \tag{9}$$

This is equivalent to assuming[11] $c = 3\times10^{10}$ cm · sec^{-1}, and introduces an error of 0.07 percent. An error of this size is by no means negligible, in the case of the work discussed in this section, and the writer should like to emphasize, this fact. Scarcely any one uses $c = 3\times10^{10}$ when reducing λ to ν, and thus in the *same* equation (8), it is quite customary to use two *different* values of c.

[7] E. O. Lawrence, Phys. Rev. **28,** 947 (1926).

[8] R. T. Birge, Science **64,** 180 (1926).

[9] E. O. Lawrence, Phys. Rev. **27,** 809 (1926), abstract 49.

[10] See for instance Ladenburg, HP. **23,** 290. On the other hand, Henning and Jaeger (HP. **2,** 511) give the correct reduction from volts to abs. *es* units.

[11] More accurately, it assumes c' (the ratio of the *es* to *em* units) $= 3\times10^{10}$. As noted in section A, the best value of $c'(=2.9979)$ agrees with the slightly more accurate value of c $(=2.99796)$, and one can therefore safely assume $c' = c$ in all calculations.

The "term" of Hg corresponding to the ordinary ionization potential is[12] 84178.5 cm^{-1}. Hence in Eq. (8) we have

$$h=\frac{(1.00046)(10.40)10^8(4.770\times10^{-10})}{(2.99796\times10^{10})^2\,84178.5}=(6.560\pm0.015)\times10^{-27} \qquad (10)$$

in which the probable error in V' is 0.2 percent, as already stated, and in e, 0.1 percent. The errors of the other factors are negligibly small.

(c) *X-ray Continuous Spectrum.* This method, like the preceding, uses Eq. (8), λ being measured by means of a calcite crystal, i.e.,

$$\lambda=2d\sin\theta \qquad (11)$$

where d is the grating space of calcite, and θ the angle at which the given wave-length shows constructive interference. Substituting Eq. (11) in Eq. (8), we obtain

$$h/e=pq\,2d(V'\sin\theta)10^8/c^2 \qquad (12)$$

In 1919 the most accurate work by this method was that of Blake and Duane.[13] Since then Duane, Palmer and Yeh[14] have carried out a still more accurate investigation, and their result was used by the writer[8] in his 1926 article. The resulting value of h, as given by these investigators, is $(6.556\pm0.009)\times10^{-27}$. Another result for which equal accuracy is claimed, is that by Wagner.[15] The full details of this latter investigation have not been published, but Ladenburg[16] gives a complete list of Wagner's experimental results. Wagner's own result, based on $c=3\times10^{10}$, is 6.526 ± 0.01. Ladenburg, using Eq. (9), with $c=2.9985\times10^{10}$, changes this to 6.529 ± 0.01.

Duane, Palmer and Yeh used a definitely known potential (in int. volts) and measured the angle θ of their ionization chamber at which the ionization suddenly started (or stopped). This gives the critical ionization frequency (or wave-length) corresponding to the given voltage. On the other hand, Wagner used a known wave-length (or rather a series of known wave-lengths) and varied the voltage for a given wave-length, until ionization suddenly began (or ceased). As shown by Eq. (12), both methods involve a knowledge of the calcite grating space d. This important constant has been discussed in section K. In that section the absolute wave-lengths of rays were used to evaluate d', the true grating space, and d' was then used with other known constants to evaluate the electronic charge e. In the present section we use the finally adopted value of e, $(4.770\pm0.005)\times10^{-10}$ abs. *es* units, with these same constants, to evaluate d'. We therefore write Eq. (10) of section K as

[12] A. Fowler, Report on Line Series, p. 149.
[13] F. C. Blake and W. Duane, Phys. Rev. **10,** 624 (1917).
[14] W. Duane, H. H. Palmer, and Chi-Sun Yeh, Proc. Nat. Acad. Sci. **7,** 237 (1921).
[15] E. Wagner, Physik. Zeits. **21,** 621 (1920).
[16] HP **23,** 296.

$$d_{20}' = \left[\frac{4.770 \times 10^{-10}}{1.7176 \times 10^{13}}\right]^{1/3} = (3.0283 \pm 0.0010) \times 10^{-8} \text{ cm}. \quad (13)$$

It should be noted that this value of d'_{20} *includes* the result of the x-ray work, since the value of e used in Eq. (13) is the weighted average obtained from both oil-drop and x-ray work. Thus we might have used $e = 4.768$ to get a value of d' based on oil-drop work. A second value of d' might then be obtained from absolute x-ray measurements. The weighted average of these two values of d' would however be just the value given in Eq. (13), provided we use the data and probable errors indicated in section K.

Using Eq. (5) of section K, we obtain for the effective grating space of calcite at 20°C, for the first order spectrum,

$$d_{20} = (3.0279 \pm 0.0010) \times 10^{-8} \text{cm}. \quad (14)$$

This value is now to be substituted in Eq. (12), together with the direct experimental value of $V' \sin\theta$. For the latter quantity Duane, Palmer and Yeh[14] found 2039.9 ± 1 int. volts, the work being performed at a mean temperature of about 20°C. Using the values of p, q, e, and c already adopted, we have

$$h = \frac{(1.00046)2(3.0279 \times 10^{-8})(2039.9)(4.770 \times 10^{-10})10^8}{(2.99796 \times 10^{10})^2} = 6.5591 \times 10^{-27} \text{ erg} \cdot \text{sec}. \quad (15)$$

with the assumed probable errors in the various factors, the final result is

$$h = (6.559 \pm 0.008) \times 10^{-27} \text{ erg} \cdot \text{sec}. \quad (16)$$

This error is due almost entirely to that in e. If the only method for determining d were by means of the density of calcite combined with the oil-drop value of e, the error in $e^{4/3}$ should be used in connection with h, the error in d due to $e^{1/3}$ being withdrawn. This has been pointed out by the writer.[1] The error in h then becomes ±0.009. With d determined in the two different ways, the error in h lies somewhere between 0.008 and 0.009.

The result found by Wagner[15] is now to be similarly revised. Using in Eq. (12) 2.99796 for c, in place of 3.00, $pq = 1.00046$ in place of unity, and $e = 4.770$ in place of 4.774, we obtain 6.532 ± 0.010, in place of 6.526 ± 0.010. It is difficult to judge what revision is required in the values of λ used by Wagner, but any such change is probably rather small. We thus have, as the two best values of h, from x-ray data, 6.559 ± 0.008 (or 0.009) and 6.532 ± 0.010. The writer does not feel competent to discuss the probable causes for this discrepancy, and these have not been discussed by others. The work of Wagner has not yet been published in sufficient detail to allow an examination of possible sources of error. For this reason in adopting a weighted average of the two results, I shall give only one-half as much weight to Wagner. Since the two results differ by much more than the probable error of

either, I use the regular least squares probable error of the resulting weighted average. Hence, from x-ray data,[16a]

$$h = (6.550 \pm 0.009) \times 10^{-27} \text{ erg} \cdot \text{sec.} \tag{17}$$

(d) *Photo-electric Effect.* The most accurate evaluation of h, from photo-electric work, is that by Lukirsky and Prilezaev.[17] These investigators use a somewhat different experimental technique from that employed earlier by Millikan,[18] and obtain a simple empirical relation for the ionization current as a function of voltage. By means of this, the actual curve may be transposed into a linear graph, making the extrapolation to zero current more certain. They also carry the readings very close to this zero point.

The equation for evaluating h is the same as that used in the previous method, except that now the energy (P) to pull an electron out of the metal is no longer negligible compared to $h\nu$. Hence we write

$$Ve = h\nu - P. \tag{18}$$

In order to eliminate P, it is customary to use light of varying frequencies, measuring for each the critical voltage V at which ionization starts. V is then plotted against ν, and the resulting curves should be linear, with a slope

$$dV/d\nu = h/e. \tag{19}$$

With V measured as V' int. volts, and ν as ν' cm^{-1}, we have

$$\frac{dV}{d\nu} = \frac{pa10^8 dV'}{c^2 d\nu'} = \frac{h}{e}. \tag{20}$$

Lukirsky and Prilezaev obtain a considerable number of such curves, using the metals Al, Zn, Sn, Ni, Cd, Cu, and Pt. Six curves, three with Zn, two with Al, and one with Ni, gave the most reliable results, and the six resulting values of h are tabulated. Unfortunately these investigators give no detailed data, and give no indication of the actual equation used for evaluating h. They show a small-sized plot of certain sample graphs. Their final value of h is 6.543×10^{-27} erg·sec, the individual results being 6.539, 6.542, 6.540, 6.556, 6.536, and 6.546. These give a probable error of only 0.035 percent, based on mutual consistency. Because of the lack of information as to the source of these values, and especially because of the surprisingly small value of the resulting error, it seems safer to assume about 0.1 percent as the true probable error of the result, exclusive of the error in the auxiliary factors.

[16a] H. Feder, Ann. d. Physik (5) **1,** 494 (1929), working under E. Wagner, has just obtained by this same method, $h = 6.547 \pm 0.003$, exclusive of the error in c. He explains Wagner's low result as directly due to the large size of the slit width used. His own value, 6.547, is identical with the value of h finally adopted in this section, and is lowered only one part in 10^4 by the use of my own values of all auxiliary constants.

[17] P. Lukirsky and S. Prilezaev, Zeits. f. Physik **49,** 236 (1928).

[18] R. A. Millikan, Phys. Rev. **7,** 355 (1916).

As just noted, there is no statement as to what values were used in Eq. (20). It seems quite certain that they used $e=4.774$, in place of our 4.770. Because of this, their final value of h should be lowered 0.0055×10^{-27}. It is very probable that they used $pq=1$, instead of our 1.00046. Their h should therefore be raised 0.0030×10^{-27}. According to the value of c used in each factor of c^2 in Eq. (20), there are other possible increases in h. My final conclusion is that the *probable* correction to their value of h lies between -0.0025 and $+0.0020$, and the *possible* correction between -0.0025 and $+0.0064$, all in 10^{-27} erg·sec units. Under the circumstances I adopt their original value, but add a probable error of 0.003×10^{-27} ($=0.05$ percent) to cover this uncertainty. The only other appreciable error is 0.1 percent due to e. The final probable error $(0.1^2+0.05^2+0.1^2)^{1/2}$ is then 0.15 percent. Hence

$$h=(6.543\pm0.010)\times10^{-27}\ \text{erg}\cdot\text{sec}. \tag{21}$$

These investigators estimate their final error in h as 0.1 to 0.2 percent, in agreement with my own assumption.

(e) *Wien's Displacement Law and the Planck Equation.* h may be determined from radiation constants in two different ways.[19] The first is by means of the constant c_2, occuring in the Wien displacement law. This law is

$$\lambda_{\max}T=c_2/\beta=A. \tag{22}$$

where $\beta=4.9651$ (root of the equation $e^{-\beta}+\beta/5-1=0$). The radiation constant c_2 occurs also in Planck's black body radiation law in the form

$$c_2=hc/k. \tag{23}$$

where $c=$ velocity of light, k(Boltzmann constant) $=R_0/N_0$, R_0 (gas constant per mole) $=\nu_n A_n/T_0$, and N_0 (Avogadro's number) $=Fc/e$. Substituting the values of ν_n, A_n, T_0, F and e previously adopted,

$$N_0=(6.0644\pm0.0061)\times10^{23}\ \text{mole}^{-1}. \tag{24}$$

$$R_0=(8.3136\pm0.0010)\times10^{7}\ \text{erg}\cdot\text{deg}^{-1}\cdot\text{mole}^{-1}. \tag{25}$$

$$k\ =(1.3709\pm0.0014)\times10^{-16}\ \text{erg}\cdot\text{deg}^{-1}. \tag{26}$$

In 1919 the writer asked Coblentz, an expert in this field, what in his opinion was then the best value of c_2. Dr. Coblentz recommended 1.433 cm·deg. and this value was adopted. In a long critical review of the radiation constants, published three years later, Coblentz[20] gives 1.432 as the most probable value. No probable error is given but the four results, obtained by four different investigators, for which 1.432 is the weighted average, are 1.436, 1.430, 1.430, and 1.4318, the last being the final result of

[19] A good description of methods for measuring the radiation constants is given by W. W. Coblentz in Glazebrook's "Dictionary of Applied Physics" IV, 541–565.

[20] W. W. Coblentz, Bur. Standards, Bull. **17**, 7 (1922). Much of the contents of this article is given also by Coblentz in J. Optical Soc. Am. **5**, 131 (1921). For a still later brief report see Coblentz, ibid **8**, 11 (1924).

Coblentz' own work in this field. Ladenburg,[21] in his HP article on h, gives 1.432 ± 0.006 as the most probable value. The separate results which he used in obtaining this average are 1.425 to 1.441, 1.4295 ± 0.007, 1.435, 1.4318, and 1.430. The chief error in all these results arises from the various corrections which must be applied to the directly observed values. Thus Coblentz' original value[22] of 1.4456 has become 1.4318 in his latest article,[20] as a result of various additional corrections. Coblentz originally considered his accuracy quite high, but in his later articles no definite estimate of the probable error is given. The writer can however see no justification for Ladenburg's assumption of 0.006 as the probable error of the weighted mean of all observations now available. Even if one grants that each separate investigation may be in error by this amount, the average surely has a smaller probable error, unless there are large, as yet unknown and unsuspected sources of *constant* error. I believe 0.003 is a much more reasonable estimate of error. Both Coblentz and Ladenburg agree on the absolute value. Hence I adopt

$$c_2 = 1.432 \pm 0.003 \text{ cm} \cdot \text{deg}. \tag{27}$$

Substituting in Eq. (23), one obtains

$$h = (6.548 \pm 0.015) \times 10^{-27} \text{ erg} \cdot \text{sec}. \tag{28}$$

It may be noted, in passing, that the radiation constant c_2 occurs in the general Boltzmann factor $e^{-\epsilon/kT}$, (ϵ = energy, T = absolute temperature) in the form $e^{-c_2\nu/T} = e^{-c_2/\lambda T}$, where ν in cm^{-1}, or λ in cm, is the quantum equivalent of ϵ ergs.

(f) *The Stefan-Boltzmann Law and the Planck Equation.* The second method for determining h by means of radiation constants is through the Stefan-Boltzmann law

$$E = \sigma T^4 = acT^4/4 \tag{29}$$

h is connected with σ, using Planck's law, by the relation

$$h = (2\pi^5 k^4 / 15 c^2 \sigma)^{1/3}, \tag{30}$$

k and c having already been defined. Just as in the case of c_2, there seems to be a difference of opinion among experts in this field concerning the accuracy with which σ may be measured. The best value, in 1919, according to the published literature, was that obtained by Coblentz,[23] namely $(5.722 \pm 0.012) \times 10^{-5}$ erg $\cdot$ cm^{-2} $\cdot$ deg^{-4} $\cdot$ sec^{-1}. In his more recent critical discussion of this question, Coblentz[20] gives a table of all available data, and concludes that the most probable value lies between 5.72 and 5.73.

[21] HP **23,** 303.

[22] W. W. Coblentz, Bur. Standards, Bull **10,** 1–77 (1914).

[23] W. W. Coblentz, Proc. Nat. Acad. Sci. **3,** 504 (1917). Full details were published later in Bur. Standards, Bull. **15,** 529 (1920). His earlier work is given in Bur. Standards, Bull. **12,** 503 (1916).

Since this 1922 article by Coblentz, there have been two new determinations of σ, one by Hoffman,[24] using the method of Westphal,[25] giving $\sigma = 5.764 \pm 0.052$, and the other by Kussman,[26] using the modified Angstrom pyrheliometer. This latter method was used also by Coblentz[23] giving 5.722 as stated, by Gerlach[27] giving 5.80, and by Kahanowicz[28] giving 5.69 to 5.73 as corrected by Coblentz.[20] Kussman obtained $\sigma = 5.795 \pm$ one percent. Ladenburg,[29] in his HP article on h, quotes merely the four results by Gerlach,[27] Hoffman,[24] Coblentz[23] and Kussman.[26] He adopts the *unweighted* average of these four results, namely 5.77, and ± 0.06 as the probable error. Here again Ladenburg assumes for the probable error of an average, a value which he himself estimates to be the probable error of each separate determination. He agrees with Gerlach[27] in thinking that Coblentz' true error is more nearly 0.06 than 0.012. It is interesting to note that the actual experimental results of Kussman[26] and Coblentz[23] are in almost perfect agreement, as shown by a plot made by Kussman. The discrepancy in their final results is due entirely to the difference in the correction applied by these two investigators for the lack of complete absorption of the receiver. Michel and Kussman[30] claim to prove that the correction Coblentz applied is too small, thus accounting for his lower value of σ.

In deciding upon the most probable value of σ there is a certain psychological factor involved. We know from the preceding discussion what is at least approximately the most probable value of h. Now the values of σ found by Kussman and by Hoffman, as well as Gerlach's earlier value of 5.80, correspond to impossibly low values of h (6.528 to 6.514), while Coblentz' own result gives an h in essential agreement with that obtained by more accurate methods.[31] This fact tends, in itself, to indicate the correctness of Coblentz' correction for incomplete absorption of the receiver, as opposed to the correction used by Kussman.

It appears to the writer that Coblentz' estimated error for his own work (5.722 ± 0.012) is too small, but that his final average of the work of all investigators up to 1922 (5.72 to 5.73) should be more trustworthy than any single determination. I shall therefore choose 5.725 as this average, and 0.02 for its probable error. This result is then to be averaged with the more recent work by Hoffman and by Kussman, using the usual rule for weighting. The final weighted average is

[24] K. Hoffman, Zeits. f. Physik **14,** 301 (1923).
[25] W. H. Westphal, Verhandl. deut. physik. Ges. **14,** 987 (1912) and **15,** 897 (1913).
[26] A. Kussman, Zeits. f. Physik **25,** 58 (1924).
[27] W. Gerlach, Ann. Physik **50,** 259 (1916) and Zeits. f. Physik **2,** 76 (1920).
[28] M. Kahanowicz, Nuovo Cimento (6) **13,** 142 (1917).
[29] HP **23,** 304.
[30] G. Michel and A. Kussman, Zeits. f. Physik **18,** 263 (1923).
[31] L. Strum, Zeits. f. Physik **51,** 287 (1928), suggests that the difference between Ladenburg's assumed experimental $\sigma = 5.77$, and the value (5.72) given by the Planck equation with the value of h assumed as known, is real and is due to the fact that the Planck equation is not strictly correct.

$$\sigma = (5.735 \pm 0.011) \times 10^{-5} \text{ erg} \cdot \text{cm}^{-2} \cdot \text{deg}^{-4} \cdot \text{sec}^{-1}. \quad (31)$$

Hence, in Eq. (29), $a = 4\sigma/c = (7.652 \pm 0.015) \times 10^{-15}$ erg $\cdot$ cm$^{-3} \cdot$ deg^{-4}

Since this weighted average value of σ was calculated and adopted, there has appeared a further determination of this quantity, by Hoare.[32] He used a Callendar radio balance, and the advantage of the method is that both the source and the receiver are essentially "black bodies," so that no corrections are necessary. This is very important for, as already noted, the chief difference of opinion regarding the value of σ has hinged on the correction to be applied for incomplete absorption. Hoare obtains $\sigma = 5.735$, agreeing exactly with the value already adopted. He gives no estimate of probable error, but lists 38 separate results, showing an average deviation from the average of only 0.016. The inclusion of this new result obviously leaves the average value unchanged, and I shall leave the probable error of Eq. (31) unchanged. Objection might have been made to this adopted error as too small, but such an objection can hardly hold in the face of Hoare's work, which seems to be the most accurate now available. This new work also speaks against Strum's assumption of an inadequacy of Planck's formula.

Using Eq. (31) in Eq (30) we obtain

$$h = (6.539 \pm 0.010) \times 10^{-27} \text{ erg} \cdot \text{sec}. \quad (32)$$

(g) *Summary* We have now obtained six determinations of h, as listed in the table.

Table of values of h

Method	h	Power of e involved
Rydberg constant	6.547 ± 0.011	5/3
Ionization potentials	6.560 ± 0.015	3/3
X-rays	6.550 ± 0.009	4/3
Photo-electric	6.543 ± 0.010	3/3
c_2	6.548 ± 0.015	3/3
σ	6.539 ± 0.010	4/3

The weighted average of these six results is $h = 6.5466 \times 10^{-27}$ erg $\cdot$ sec The least squares probable error of the average, determined by the usual formula, is ± 0.0017, an error definitely smaller than that deduced in 1919, indicating an increased consistency of the data. It may seem surprising that the probable error of each of the six results is so nearly the same. This is due to the fact that in each case, the error listed is due chiefly to the error in e. Each result involves e to the power given in the last column of the table. The *weighted* average power is 1.26. Hence the weighted average value of h can be assumed to involve e to the power 1.26. The error in h due to this is $1.26 \times 10^{-3} \times 6.55 = 0.00825$. Combining this with the above least squares error in h, we have a total final error of 0.00842, essentially no greater than the error due to e alone. Hence our final result is

[32] F. E. Hoare, Phil. Mag. **6**, 828 (1928).

$$h = (6.547 \pm 0.008) \times 10^{-27} \text{ erg} \cdot \text{sec.} \tag{33}$$

It is of interest to note that this value of h is identical with Ladenburg's most recent estimate.[3] This identity is however spurious, since Ladenburg assumes $e = 4.774 \times 10^{-10}$. If this older value of e had been used in the present work, we should have obtained $h = 6.5535$, in practically exact agreement with the writer's 1919 value (6.5543).

Another potentially accurate method for evaluating h is given by the Compton shift of x-ray lines. The theoretical equation for this is

$$\Delta\lambda = (h/mc)(1 - \cos\phi) \tag{34}$$

where m is the mass of an electron, as deduced from the values of e and e/m. Since, however, h varies in value with e, Eq. (34) can better be used to evaluate e/m, as mentioned incidentally in section L. We can in fact write

$$\Delta\lambda = (h/e)(e/m)(1 - \cos\phi) \tag{35}$$

in which e as usual is in *es* units, and e/m in *em* units. Then

$$\frac{e}{m} = \frac{\Delta\lambda}{(h/e)(1 - \cos\phi)}$$

The most accurate work on this subject has been done by Sharp,[33] who obtains $\Delta\lambda = (0.04825 \pm 0.00017) \times 10^{-8}$ cm, for $(1 - \cos\phi) = (1.984 \pm 0.001)$. With the adopted values of h and e, we have[34] $h/e = (1.3725 \pm 0.0005) \times 10^{-17}$ erg $\cdot$ sec $\cdot$ es^{-1}. Substituting in Eq. (36) one finds $e/m = (1.772 \pm 0.006) \times 10^{7}$ abs. *em* units, the final error being due almost entirely to the error in $\Delta\lambda$. It seems to the writer possibly significant that this value agrees better with the deflection than with the spectroscopic value of e/m, for the theory used in the derivation of Eq. (35) is essentially the collision theory of classical dynamics for free electrons. Greater accuracy is however needed before any real conclusion can be drawn.[35]

Section N

Tables

The various important constants and ratios which have been evaluated in sections A to M are collected in Table a of this section. In the process of evaluating these constants, it has been necessary to calculate certain auxiliary constants, and also to use certain conventional quantities, such as g_{45} and g_n. All such quantities are listed in Table b, in the order in which they appear in the preceding sections. The section in which each quantity

[33] H. M. Sharp, Phys. Rev. **26**, 691 (1925).

[34] See footnote 2 of section N.

[35] B. Davis and D. P. Mitchell, Phys. Rev. **32**, 331 (1928), have recently found, for the "undisplaced" line, a fine structure corresponding directly to the *Raman effect* observed with ordinary light. The observed displacement of the lines is considerably less than the Compton shift, and so the accuracy of Sharp's work is presumably unaffected by this new phenomenon

is discussed is indicated on the left side of the table. For convenience the indirect calculation of R_∞, $R_\infty \cdot c$, e/m, c_2, σ and a are also included in this table. All of these indirect calculations depend on the finally adopted value of h, and so could not be obtained when the more directly observed value was under consideration. It should be noted that in the case of R_∞, $R_\infty \cdot c$ and e/m, the direct evaluation is the more accurate, while in the case of c_2, σ and a the indirect evaluation is the more accurate.

Table a. Principal Constants and Ratios

Section		
A	Velocity of light	$c = (2.99796 \pm 0.00004) \times 10^{10}$ cm·sec^{-1}
B	Gravitation constant	$G = (6.664 \pm 0.002) \times 10^{-8}$ dyne·cm^2·g^{-2}
C	Liter	$l = 1000.027 \pm 0.001$ cm^3
D	Volume of perfect gas (0°C, A_n)	$v_n = (22.4141 \pm 0.0008) \times 10^3$ cm^3·mole^{-1}
		$R_n = 22.4135 \pm 0.0008$ liter·mole^{-1}
D	Volume of perfect gas (0°C, A_{45})	$R = 22.4146 \pm 0.0008$ liter·mole^{-1}
E	International ohm (= p abs-ohm)	$p = 1.00051 \pm 0.00002$
E	International ampere (= q abs-amp)	$q = 0.99995 \pm 0.00005$
F	Atomic weights	
	O (oxygen) = 16.0000 (definition)	Ag (silver) = 107.880 ± 0.001
	H (hydrogen) = 1.00777 ± 0.00002	I (iodine) = 126.932 ± 0.002
	He (helium) = 4.0022 ± 0.0004	C (carbon) = 12.003 ± 0.001
	N (nitrogen) = 14.0083 ± 0.0008	Ca (calcium) = 40.075 ± 0.005
G	Normal atmosphere	$A_n = (1.013249 \pm 0.000003) \times 10^6$ dyne·cm^{-2}
G	45° atmosphere	$A_{45} = (1.013199 \pm 0.000003) \times 10^6$ dyne·cm^{-2}
H	Ice point (absolute scale)	$T_0 = 273.18 \pm 0.03$°K
I	Mechanical equivalent of heat[1]	$J_{15} = 4.1852 \pm 0.0006$ abs-joule·cal$_{15}^{-1}$
		$J_{20} = 4.1813 \pm 0.0006$ abs-joule·cal$_{20}^{-1}$
I	Electrical equivalent of heat[1]	$J'_{15} = 4.1835 \pm 0.0007$ int-joule·cal$_{15}^{-1}$
		$J'_{20} = 4.1796 \pm 0.0007$ int-joule·cal$_{20}^{-1}$
J	Faraday constant[1]	$F = 96494 \pm 5$ int-coul·g-equiv^{-1}
		$= 96489 \pm 7$ abs-coul·g-equiv^{-1}
		$= 9648.9 \pm 0.7$ abs-*em*-unit·g-equiv^{-1}
		$Fc = (2.8927_0 \pm 0.0002) \times 10^{14}$ abs-*es*-unit·g-equiv^{-1}
K	Electronic charge[1]	$e = (4.770 \pm 0.005) \times 10^{-10}$ abs-*es*-units
		$e/c = (1.5910_3 \pm 0.0016) \times 10^{-20}$ abs-*em*-units
L	Specific electronic charge (spectroscopic)[1]	$e/m = (1.761 \pm 0.001) \times 10^7$ abs-*em*-unit·g^{-1}
		$(e/m)c = (5.279_{41} \pm 0.003) \times 10^{17}$ abs-*es*-unit·g^{-1}
L	Specific electronic charge (deflection)[1]	$e/m = (1.769 \pm 0.002) \times 10^7$ abs-*em*-unit·g^{-1}
		$(e/m)c = (5.303_{39} \pm 0.006) \times 10^{17}$ abs-*es*-unit·g^{-1}
M	Planck constant	$h = (6.547 \pm 0.008) \times 10^{-27}$ erg·sec

In addition to certain constants listed in Table b, there are many others which are functions of constants given in Tables a and b. A number of these derived constants are collected in Table c. This table could be extended almost indefinitely, but an attempt has been made to include the more important or more frequently used constants. The process for obtaining the correct probable error for many of the constants of Tables b and c is sometimes involved. The general principle underlying it is stated and illustrated in footnote 2 of this section. The possible theoretical significance of certain relations appearing in these tables is considered in section O.

[1] In the expressions for all derived quantities in these tables, e is to be understood as in abs-*es*-units, (e/m) as in abs-*em*-units, R_H, R_{He} and R_∞ as in cm^{-1} units, and J as in joules per cal. F is in abs.-*em*-units except as noted in one constant of Table c.

Table b. Additional Quantities Evaluated or Used in Connection with Table a.

Section		
A	Ratio of *es* to *em* units (direct)	$c' = (2.9971 \pm 0.0001) \times 10^{10}$ cm$^{1/2}\cdot$sec$^{-1/2}\cdot$ int $\Omega^{-1/2}$ $= (2.9979 \pm 0.0001) \times 10^{10}$ cm$\cdot$sec^{-1}
A	Ratio of *es* to *em* units (indirect)	$c' = c = (2.99796 \pm 0.00004) \times 10^{10}$ cm$\cdot$sec^{-1}
B	Acceleration of gravity (45°)	$g_{45} = 980.616$ cm$\cdot$sec^{-2}
B	Mean density of earth	$\delta = 5.522 \pm 0.002$ g$\cdot$cm^{-3}
C	Maximum density of water	$\delta_m = 0.999973 \pm 0.000001$ g$\cdot$cm^{-3}
D	Acceleration of gravity (normal)	$g_n = 980.665$ cm$\cdot$sec^{-2}
D	Density of oxygen gas (0°C, A_{45})	$L = 1.428965 \pm 0.000030$ g$\cdot$liter^{-1}
D	Factor converting oxygen (0°C, A_{45}) to ideal gas	$1-\alpha = 1.000927 \pm 0.000030$
E	International coulomb (= q abs-coul)	$q = 0.99995 \pm 0.00005$
E	International henry (= p abs-henry)	$p = 1.00051 \pm 0.00002$
E	International volt (= pq abs-volt)	$pq = 1.00046 \pm 0.00005$
E	International joule (= pq^2 abs-joule)	$pq^2 = 1.00041 \pm 0.00010$
E	International gauss (= q abs-gauss)	
F	Density of nitrogen (0°C, A_{45})	$L = 1.25046 \pm 0.000045$ g$\cdot$liter^{-1}
F	Factor converting nitrogen (0°C, A_{45}) to ideal gas	$1-\alpha = 1.00043 \pm 0.00002$
G	Specific gravity of Hg (0°C, A_n) referred to air-free water at maximum density	$\rho_n = 13.59546 \pm 0.00003$
G	Density of Hg (0°C, A_n)	$D_n = 13.59509 \pm 0.00003$ g$\cdot$cm^{-3}
J	Electrochemical equivalent of silver	$E_{Ag} = (1.11800 \pm 0.00005) \times 10^{-3}$ g$\cdot$int-coul^{-1} $= (1.11805 \pm 0.00007) \times 10^{-3}$ g$\cdot$abs-coul^{-1}
K	Arbitrary calcite grating space (18°C)	$d_{18}'' = 3.02904 \times 10^{-8}$ cm
K	Density of calcite (20°C)	$\rho = 2.7102 \pm 0.0004$ g$\cdot$cm^{-3}
K	Structural constant of calcite (20°C)	$\phi(\beta) = 1.09630 \pm 0.00007$
L	Rydberg constant for hydrogen[1]	$R_H = 109677.759 \pm 0.05$ cm^{-1}
L	Rydberg constant for ionized helium[1]	$R_{He} = 109722.403 \pm 0.05$ cm^{-1}
L	Wave-length of red Cd line (15°C, A_n)	λ(Cd) = 6438.4696 I.A. (definition of I.A. unit)
M	Rydberg constant for infinite mass[1]	$R_\infty = 109737.42 \pm 0.06$ cm^{-1} $R_\infty c = (3.28988 \pm 0.00004) \times 10^{15}$ sec^{-1}
	R_∞(indirect)[1,2]	$= 2\pi^2 e^5/h^3 c^2 (e/m)_{sp} = 10974_4 \pm 170$ cm^{-1}
		$R_\infty c$ (indirect)[1,2] $= (3.290_{03} \pm 0.005) \times 10^{15}$ sec^{-1}
M	e/m (spectroscopic, indirect)[1,2]	$e/m = 2\pi^2 e^5/h^3 c^2 R_\infty = (1.7611_2 \pm 0.0025) \times 10^7$ abs-*em*-unit$\cdot$g^{-1}
M	True grating space of calcite (20°C)	$d'_{20} = (3.0283 \pm 0.0010) \times 10^{-8}$ cm

[2] The error is derived by the process used in section M, in getting the final probable error in h itself. This process is used in these tables for every constant which involves e in more than one factor. It may be illustrated as follows. Let $y(=10^{-3})$ denote the proportional error in e, and $z(=0.0017/6.547=0.26\times10^{-3})$ the proportional error in h excluding the error in e. Since h varies as $e^{1.26}$, the final proportional error in h is $[(1.26y)^2+z^2]^{1/2}=1.287\times10^{-3}$, corresponding to the absolute error 0.00842×10^{-27} erg$\cdot$sec, already given in section M. Similarly, since h/e varies as $e^{0.26}$, the proportional error in h/e is given by $[(0.26y)^2+z^2]^{1/2}=0.368\times10^{-3}$. As a third example consider e^5/h^3, as it occurs in the indirect evaluation of R_∞ or e/m. This ratio varies as $e^{1.22}$. The proportional error in e^5/h^3 is then $[(1.22y)^2+(3z)^2]^{1/2}=1.45\times10^{-3}$. A fourth example is the indirect calculation of $c_2(=hc/k)$. Since k varies directly as e, h/k varies as $e^{0.26}$. Let u be the proportional error in k, exclusive of the error due to e, $(u=1.27\times10^{-4})$. Then the error in h/k is $[(0.26y)^2+z^2+u^2]^{1/2}=3.89\times10^{-4}$, the error in c being negligible. As a final example consider the indirect calculation of σ, in Table b. The error here is $[(0.22y)^2+(4u)^2+(3z)^2]^{1/2}=9.6\times10^{-4}$ In the case of some of the derived constants, notably the chemical constants S_0 and i_0 of Table c, the calculation of the error is much more involved. This matter is mentioned again in Section O.

M Effective grating space of calcite (20°C) ..... $d_{20} = (3.0279 \pm 0.0010) \times 10^{-8}$ cm

M Avogadro's number[1] ..... $N_0 = Fc/e = (6.064_{36} \pm 0.006) \times 10^{23}$ mole^{-1}

M Gas constant per mole .. $R_0 = \nu_n A_n / T_0 = (8.3136_0 \pm 0.0010) \times 10^7$ erg·deg^{-1}·mole^{-1}

$R_0' = R_0/(J_{15} \times 10^7) = 1.9864_3 \pm 0.0004$ cal$_{15}$·deg^{-1}·mole^{-1}

$R_0'' = R_0/(A_n \times 1000.027) = (8.2046_4 \pm 0.0009) \times 10^{-2}$ liter·atmos·deg^{-1}·mole^{-1}

$R_0''' = R_0/A_n = \nu_n/T_0 = 82.048_8 \pm 0.009$ cm^3·atmos·deg^{-1}·mole^{-1}

M Boltzmann constant ..... $k = R_0/N_0 = (1.3708_9 \pm 0.0014) \times 10^{-16}$ erg·deg^{-1}

M Second radiation constant (experimental) ..... $c_2 = 1.432 \pm 0.003$ cm·deg

M Second radiation constant (indirect)[2] ..... $c_2 = hc/k = 1.4317_4 \pm 0.0006$ cm·deg

M Stefan-Boltzmann constant (experimental) ..... $\sigma = (5.735 \pm 0.011) \times 10^{-5}$ erg·cm^{-2}·deg^{-4}·sec^{-1}

M Stefan-Boltzmann constant (indirect)[2] ..... $\sigma = 2\pi^5 k^4/15c^2h^3 = (5.713_9 \pm 0.006) \times 10^{-5}$ erg·cm^{-2}·deg^{-4}·sec^{-1}

M Radiation density constant (experimental) ..... $a = 4\sigma/c = (7.651_8 \pm 0.015) \times 10^{-15}$ erg·cm^{-3}·deg^{-4}

M Radiation density constant (indirect)[2] ..... $a = 8\pi^5 k^4/15c^3h^3 = (7.623_7 \pm 0.007) \times 10^{-15}$ erg·cm^{-3}·deg^{-4}

As in the case of the basic constants listed on page 17 of Vol. 1 of the ICT, the constants of the present Table a are, for purposes of computation, to be taken as exactly correct. That is, all additional digits in each constant are to be assumed as zero. The real probable error in each constant is of course that indicated in the table, and each such constant has an accepted value carried only to the number of significant figures required by the adopted probable error. On the other hand, the various *derived* constants of Table c (and the occasional derived constant appearing in Tables a and b) are given with one and often two more digits than is required by the probable error. Such additional digits are printed below the line, and have been added in order that calculations of derived constants, when made in different ways (often in terms of other derived constants) shall not introduce any appreciable error. The ICT accomplishes the same result by printing the logarithm, to seven digits, of each constant. This does not help in calculations of derived constants, made with a calculating machine. If such a machine is not available, the writer has found Crelle's multiplication tables more convenient than logarithms.

A brief explanation of certain points of nomenclature seems desirable.

(a) The nomenclature adopted in this paper is intended to conform with current practise. On the other hand, an attempt has been made to avoid duplication of the same symbol, especially in the same general field. Some duplication seems unavoidable, examples of this being the use of H for atomic weight of hydrogen and for the strength of a magnetic field, σ for the Stefan-Boltzmann radiation constant and also in the expression for the rotational specific heat constant in Table c.

(b) In certain cases the writer has arbitrarily chosen a symbol for a certain constant (such as $\nu_0 = e \cdot 10^8/hc^2$, in Table c), when this constant is important and is needed in other relations. Thus the ionization potential of hydrogen equals R_H/ν_0, in Table c.

(c) e/m always indicates merely the *ratio* of charge to mass for an electron, in *em* units, this being the simplest and most commonly used

Table c. Miscellaneous Derived Constants

$c^2 = (8.98776 \pm 0.00024) \times 10^{20}$ cm$^2 \cdot$ sec^{-2}	$h/c = (2.183_{51} \pm 0.003) \times 10^{-37}$ g$\cdot$cm
$c^3 = (26.9449 \pm 0.0011) \times 10^{30}$ cm$^3 \cdot$ sec^{-3}	$e^2 = (22.752_9 \pm 0.045) \times 10^{-20}$ *es*-units^2
$c^4 = (80.7798 \pm 0.0043) \times 10^{40}$ cm$^4 \cdot$ sec^{-4}	$e^3 = (108.53_{15} \pm 0.33) \times 10^{-30}$ *es*-units^3
$h^2 = (42.86_{42} \pm 0.10) \times 10^{-54}$ erg$^2 \cdot$ sec^2	$e^4 = (517.6_{94} \pm 2.1) \times 10^{-40}$ *es*-units^4
$h^3 = (280.6_{25} \pm 1.0) \times 10^{-81}$ erg$^3 \cdot$ sec^3	$e^5 = (2469._{40} \pm 12) \times 10^{-50}$ *es*-units^5

Mass of electron (spectroscopic)	$m_0 = \dfrac{e}{c(e/m)_{sp}} = (9.035_{10} \pm 0.010) \times 10^{-28}$ g
Mass of electron (deflection)	$m_0 = \dfrac{e}{c(e/m)_{defl}} = (8.994_{25} \pm 0.014) \times 10^{-28}$ g
Atomic weight of electron (spectroscopic)	$m = F/(e/m)_{sp} = (5.479_{22} \pm 0.003) \times 10^{-4}$
Atomic weight of electron (deflection)	$m = F/(e/m)_{defl} = (5.454_{44} \pm 0.006) \times 10^{-4}$
Mass of atom of unit atomic weight	$M_0 = 1/N_0 = (1.6489_3 \pm 0.0016) \times 10^{-24}$ g
Mass of hydrogen atom	$M_H = H/N_0 = (1.6617_9 \pm 0.0017) \times 10^{-24}$ g
Atoms per gram of hydrogen	$1/M_H = (6.017_{61} \pm 0.006) \times 10^{23}$ g^{-1}
Mass of proton	$M_P = (H - m)/N_0 = (1.6608_9 \pm 0.0017) \times 10^{-24}$ g
Mass of α-particle	$M_\alpha = (He - 2m)/N_0 = (6.597_{74} \pm 0.007) \times 10^{-24}$ g
Charge in electrolysis of 1 g hydrogen	$e/M_H = F/H = 9574.5_1 \pm 0.7$ abs-*em*-units$\cdot$g^{-1}
Specific charge of proton	$e/M_P = F/(H - m) = 9579.7_3 \pm 0.7$ abs-*em*-units$\cdot$g^{-1}
Specific charge of α-particle	$2e/M_\alpha = 2F/(He - 2m) = 4823.1_1 \pm 0.6$ abs-*em*-units$\cdot$g^{-1}
Ratio mass H atom to mass electron (spectroscopic)	$\dfrac{(e/m)_{sp}}{e/M_H} = 1839._{26} \pm 1$
Ratio mass H atom to mass electron (deflection)	$\dfrac{(e/m)_{defl}}{e/M_H} = 1847._{61} \pm 2$
Ratio mass proton to mass electron (spectroscopic)	$= 1839.26 - 1 = 1838._{26} \pm 1$
Ratio mass proton to mass electron (deflection)	$= 1847.61 - 1 = 1846._{61} \pm 2$
Energy associated with unit wave-number	$\epsilon/\nu' = hc = (1.9627_{64} \pm 0.0025) \times 10^{-16}$ erg$\cdot$cm
Potential associated with unit frequency[2]	$V/\nu = h/e = (1.3725_4 \pm 0.0005) \times 10^{-17}$ erg$\cdot$sec$\cdot$*es*-unit^{-1}
Frequency associated with unit potential	$\nu/V = e/h = (7.2857_7 \pm 0.0027) \times 10^{16}$ *es*-unit$\cdot$erg$^{-1}\cdot$sec^{-1}
	$\nu/V'' = e \cdot 10^8/hc = (2.4302_5 \pm 0.0009) \times 10^{14}$ sec$^{-1}\cdot$abs-volt^{-1}
Wave-number associated with 1 abs-volt	$\nu_0 = \nu'/V'' = e \cdot 10^8/hc^2 = 8106._{31} \pm 3$ cm$^{-1}\cdot$abs-volt^{-1}
Wave-length associated with 1 abs-volt	$\lambda_0 = \lambda V'' = hc^2/e = (12336._1 \pm 5) \times 10^{-8}$ cm$\cdot$abs-volt
Energy of 1-abs-volt-electron	$h\nu/V'' = e \cdot 10^8/c = (1.5910_3 \pm 0.0016) \times 10^{-12}$ erg
Speed of 1-abs-volt-electron (spectroscopic)	$v_e = [2 \cdot 10^8 (e/m)_{sp}]^{1/2} = (5.9346_4 \pm 0.0017) \times 10^7$ cm$\cdot$sec^{-1}
Speed of 1-abs-volt-electron (deflection)	$v_e = [2 \cdot 10^8 (e/m)_{defl}]^{1/2} = (5.9481_1 \pm 0.0034) \times 10^7$ cm$\cdot$sec^{-1}

TABLE C. *Miscellaneous Derived Constants—(Continued)*

Fine structure constant	$\alpha = 2\pi e^2/hc = (7.283_{64} \pm 0.006) \times 10^{-3}$
Reciprocal of fine structure constant	$1/\alpha = 137.29_4 \pm 0.11$
Square of fine structure constant	$\alpha^2 = (5.305_{12} \pm 0.008) \times 10^{-5}$
Bohr unit of angular momentum	$h/2\pi = (1.0419_9 \pm 0.0013) \times 10^{-27}$ erg·sec
	$e^2/c = \alpha h/2\pi = (7.589_{43} \pm 0.015) \times 10^{-30}$ *es*-unit^2·sec·cm^{-1}
Magnetic moment of 1 Bohr magneton (spectroscopic)	$\mu_1 = (h/4\pi)(e/m)_{sp} = (0.9174_{70} \pm 0.0013) \times 10^{-20}$ erg·gauss^{-1}
Magnetic moment of 1 Bohr magneton (deflection)	$\mu_1 = (h/4\pi)(e/m)_{defl} = (0.9216_{38} \pm 0.0016) \times 10^{-20}$ erg·gauss^{-1}
Ratio Bohr magneton to Bohr mechanical unit (spectroscopic)	$\frac{\mu_1}{h/2\pi} = \frac{1}{2}\left(\frac{e}{m}\right)_{sp} = 0.8805 \pm 0.0005) \times 10^7$ gauss^{-1}·sec^{-1}
Ratio Bohr magneton to Bohr mechanical unit (deflection)	$\frac{\mu_1}{h/2\pi} = \frac{1}{2}\left(\frac{e}{m}\right)_{defl} = (0.8845 \pm 0.0010) \times 10^7$ gauss^{-1}·sec^{-1}
Magnetic moment per mole for one Bohr magneton per molecule (spectroscopic)	$\mu_1 N_0 = 5563._{87} \pm 10$ erg·gauss^{-1}·mole^{-1}
Same (deflection)	$\mu_1 N_0 = 5589._{14} \pm 11$ erg·gauss^{-1}·mole^{-1}
Zeeman displacement per gauss	$\frac{\Delta\nu'}{H} = \frac{(e/m)_{sp}}{4\pi c} = (4.674_{38} \pm 0.003) \times 10^{-5}$ cm^{-1}·gauss^{-1}
Band spectra constant connecting wave-number and moment of inertia, I	$h/8\pi^2 c = (27.65_{52} \pm 0.04) \times 10^{-40}$ g·cm
Same, connecting frequency and I	$h/8\pi^2 = (8.291_{87} \pm 0.011) \times 10^{-29}$ erg·sec
Same, connecting ergs and I	$h^2/8\pi^2 = (5.428_{69} \pm 0.014) \times 10^{-55}$ erg^2·sec^2
Atomic specific heat constant	$c_2/c = h/k = (4.7757_3 \pm 0.0019) \times 10^{-11}$ sec·deg
Rotational specific heat constant	$\sigma IT = h^2/8\pi^2 k = (39.59_{97} \pm 0.06) \times 10^{-40}$ g·cm^2·deg
Reduced mass of H atom	$\mu_H = R_H(m_0)_{sp}/R_\infty = (9.030_{19} \pm 0.010) \times 10^{-28}$ g
Schroedinger constant for H atom	$8\pi^2\mu_H/h^2 = (1.663_{42} \pm 0.003) \times 10^{27}$ g·erg^{-2}·sec^{-2}
Schroedinger constant for electron	$8\pi^2(m_0)_{sp}/h^2 = (1.664_{32} \pm 0.003) \times 10^{27}$ g·erg^{-2}·sec^{-2}
Ionization potential for H atom	$R_H/\nu_0 = 13.529_9 \pm 0.005$ abs-volt
Ionization potential for He^+	$4R_{He}/\nu_0 = 54.141_7 \pm 0.020$ abs-volt
Separation of nucleus and electron in normal H atom, using experimental value of R_H	$a_0 = \frac{\alpha(1-\alpha^2)^{1/2}}{4\pi R_H} = (0.5284_{56} \pm 0.0004) \times 10^{-8}$ cm
Same, using calculated value of R_H	$a_0 = \frac{h^2(1-\alpha^2)^{1/2}}{4\pi^2 e^2 \mu_H} = (0.5284_{21} \pm 0.0005) \times 10^{-8}$ cm
Radius of Bohr orbit in normal H, referred to center of mass, using experimental R_∞	$a_0' = \frac{\alpha(1-\alpha^2)^{1/2}}{4\pi R_\infty} = (0.5284_{69} \pm 0.0004) \times 10^{-8}$ cm

TABLE C. *Miscellaneous Derived Constants— (Continued)*

Same, using calculated R_∞	$a_0' = \dfrac{h^2 c (e/m)_{sp}(1-\alpha^2)^{1/2}}{4\pi^2 e^3} = (0.5281_{34} \pm 0.0005) \times 10^{-8}$ cm
Speed of electron in normal H orbit, referred to nucleus	$v_0' = \alpha c = (2.1836_1 \pm 0.0017) \times 10^8$ cm·sec^{-1}
Same, referred to center of mass, using experimental R_H and R_∞	$v_0 = \alpha c R_H / R_\infty = (2.1824_2 \pm 0.0017) \times 10^8$ cm·sec^{-1}
General spectroscopic doublet constant	$R_\infty \alpha^2 = 5.821_{71} \pm 0.009$ cm^{-1}
Hydrogen doublet constant	$\Delta\nu_H = R_H \alpha^2/16 = 0.3636_{59} \pm 0.0006$ cm^{-1}
Dispersion constant	$e^2/2\pi m_0 = ec(e/m)_{sp}/2\pi = (4.007_{97} \pm 0.005) \times 10^7$ cm^3·sec^{-2}
X-ray scattering coefficient	$\dfrac{8\pi e^4}{3m_0^2 c^4} = \dfrac{8\pi e^2 (e/m)^2_{sp}}{3c^2} = (6.576_{91} \pm 0.015) \times 10^{-25}$ cm^2
Compton shift at 90° (spectroscopic)	$\dfrac{h}{m_0 c} = \dfrac{h}{e}\left(\dfrac{e}{m}\right)_{sp} = (0.024170_4 \pm 0.000016) \times 10^{-8}$ cm
Compton shift at 90° (deflection)	$\dfrac{h}{m_0 c} = \dfrac{h}{e}\left(\dfrac{e}{m}\right)_{defl} = (0.02428_{02} \pm 0.00003) \times 10^{-8}$ cm
Wave mechanics λ of 1-abs-volt-electron (spectroscopic)	$\dfrac{h}{m_0 (v_e)_{sp}} = (12.210_0 \pm 0.006) \times 10^{-8}$ cm
Loschmidt number (0°C, A_n)	$n_0 = N_0/\nu_n = (2.705_{60} \pm 0.003) \times 10^{19}$ cm^{-3}
Wien's displacement constant (indirect)	$A = c_2/4.9651 = 0.28836_1 \pm 0.00011$ cm·deg
First radiation constant[3]	$c_1 = 8\pi hc = (4.932_{96} \pm 0.006) \times 10^{-15}$ erg·cm
	or $hc^2 = (0.5884_{29} \pm 0.0008) \times 10^{-5}$ erg·cm^2·sec^{-1}
	or $2\pi hc^2 = (3.697_{20} \pm 0.005) \times 10^{-5}$ erg·cm^2·sec^{-1}
Energy per mole for 1-abs-volt-electron per molecule	$\dfrac{F \text{ (abs. coul per gram equiv.)}}{J_{15} \text{ (abs. joules per cal.)}} = 23054._8 \pm 4$ cal$_{15}$·mole^{-1}
Translational energy of molecule at 0°C	$E_0 = 3kT_0/2 = (5.617_{49} \pm 0.006) \times 10^{-14}$ erg
Translational energy per degree	$e_0 = E_0/T_0 = 3k/2 = (2.0563_3 \pm 0.0021) \times 10^{-16}$ erg·deg^{-1}

[3] Planck's radiation law is

$$E_\lambda = \frac{c_1}{\lambda^5}\left(e^{-c_2/\lambda T} - 1\right)^{-1}.$$

E_λ may be defined in various ways, and c_1 varies accordingly. If $E_\lambda d\lambda$ denotes the energy density of unpolarized radiation in range $d\lambda$, $c_1 = 8\pi hc$ (used by Coblentz and others). If $E_\lambda d\lambda$ denotes the intensity of emission of linearly polarized radiation in range $d\lambda$, perpendicular to a surface, per unit of surface, per unit solid angle, $c_1 = hc^2$ (Planck's own nomenclature, used generally in German texts). If $E_\lambda d\lambda$ denotes the emission of unpolarized radiation in range $d\lambda$, per unit surface, in all directions (2π solid angle), $c_1 = 2\pi hc^2$ (used in this ICT and elsewhere)

TABLE C. *Miscellaneous Derived Constants—(Continued)*

Sackur-Tetrode constant[4] (ϵ, base of ln = 2.71828)	$S_0 = R_0' \ln \frac{(2\pi k)^{3/2}\epsilon^{5/2}}{h^3 N_0^{5/2}} = -11.0533_4 \pm 0.0026$ cal$_{15}\cdot$deg$^{-1}\cdot$mole^{-1}, or entropy units
Sackur-Tetrode constant, alternative form	$S_0/R_0' = -5.5643_9 \pm 0.0009$
	$S_0' (= S_0 + R_0' \ln R_0''') = -2.2985_2 \pm 0.0018$ cal$_{15}\cdot$deg$^{-1}\cdot$mole^{-1}
Chemical constant (unit at. wt., pressure in dynes$\cdot$cm^{-2})	$i_0 = \frac{3}{2}\log\frac{2\pi k^{5/3}}{N_0 h^2} = +4.4174_6 \pm 0.0004$
Same (pressure in atmospheres)	$i_0' = i_0 - \log A_n = -1.5882_5 \pm 0.0004$
Debye-Hückel constant for electrolytes[5]	$(\pi e^6/k^3)^{1/2} = (2.512_{51} \pm 0.004) \times 10^5$ *es*-unit$^3\cdot$erg$^{-3/2}\cdot$deg$^{3/2}$
Multiplier of (Curie constant)$^{1/2}$ to give magnetic moment per molecule	$(3k/N_0)^{1/2} = (2.604_{17} \pm 0.003) \times 10^{-20}$ erg$^{1/2}\cdot$deg$^{-1/2}\cdot$mole$^{1/2}$
Same, to give Bohr magnetons per molecule	$\frac{(3k/N_0)^{1/2}}{\mu_1} = 2.8384_2 \pm 0.0019$ erg$^{-1/2}\cdot$gauss$\cdot$deg$^{-1/2}\cdot$mole$^{1/2}$

[4] S_0 is to be used in the formula for the entropy of a perfect monatomic gas $S = (3/2)\ R_0' \ln M + (3/2)\ R_0' \ln T + R_0' \ln V + S_0$ where M = molecular weight in grams, T = °C abs., V = vol. in cm^3. S_0' is to be used in $S = (3/2)\ R_0' \ln M + (5/2)\ R_0' \ln T - R_0' \ln P + S_0'$ where P = pressure in normal atmospheres (A_n).

[5] See page 328 of Eucken, Jette and Lamer "Physical Chemistry."

nomenclature. On the other hand e, standing alone, indicates electronic charge in *es* units, and m_0 indicates electronic mass while m equals the atomic weight of an electron. Hence a more logical but less convenient nomenclature would have been (e/m_0) *es* units, and possibly (e'/m_0) *em* units.

(d) In the fundamental quantum relation connecting spectral frequency and energy, $\epsilon = h\nu = eV$, each side of the equation represents energy in ergs, provided all quantities are in abs. c.g.s. units. ν/V $(=e/h)$ then measures the frequency in sec^{-1} associated with one abs. *es* unit of potential. It is usually convenient to substitute the wave number (ν') or the wave-length (λ) in place of ν, and to substitute the number of abs. volts (V'') in place of V. $(V' = \text{int. volts}$, throughout this paper). The values of the various ratios, such as ν'/V'' etc., are given in Table c.

(e) An electron which has fallen through one abs. volt of potential is termed an abs. volt-electron, and its energy in ergs and speed in $cm \cdot sec^{-1}$ are given in Table c. Corresponding to any *ionization potential* of an atom or molecule in volts (V''), there is an *energy of ionization* (eV'') which can conveniently be measured in units equal to the energy of a volt-electron, and is so designated. Thus an ionization potential of 10 volts corresponds to an energy of ionization of 10 volt-electrons. Similarly it is convenient, in the case of molecules, to speak of a *dissociation potential* of, let us say, 10 volts, and a corresponding energy of dissociation (heat of dissociation) of 10 volt-electrons per molecule. The factor by which this last quantity must be multiplied to give the heat of dissociation in calories per mole is given in Table c. Unfortunately there has arisen the practise, to which the writer also must plead guilty, of designating the *heat* of dissociation as 10 volts, instead of stating, more correctly, that the equivalent dissociation *potential* is 10 volts, or that the heat of dissociation per molecule is 10 volt-electrons.

(f) The name of the units given in the tables conforms as far as possible with current practise. Certain difficulties arise in the case especially of magnetic and electrical quantities. These difficulties are connected with the unknown dimensions of magnetic permeability μ, and specific inductive capacity ϵ. It is customary in the ICT, and in most other tables and texts, to indicate these unknown dimensions by the symbols μ and ϵ. A given unit, such as the gauss, is applied only to quantities of a given set of dimensions, *including* μ and ϵ. In the present paper we are concerned only with numerical magnitudes and no particular attention has accordingly been paid to this matter of dimensions. Thus in section E, the statement that the absolute *em* unit of resistence is one $cm \cdot sec^{-1}$ involves the assumption not only of unit permeability, but also of dimensionless permeability. In a number of the equations given in Table c, the two sides of the equation do *not* check dimensionally unless one assumes μ and ϵ to be dimensionless. It follows from this that the name of the unit stated in the table applies strictly only to one side of such an equation. In such cases the unit applies to the left side of the equation, since this is the quantity being evaluated.

The right side gives merely the most direct derivation of the numerical magnitude, in terms of quantities already evaluated. Since this ambiguity does not affect the numerical magnitude, it is inconsequential in the present discussion. As examples of this situation we cite the fine structure constant α, which is dimensionless. To satisfy this condition one should write $\alpha = 2\pi e^2/\epsilon_0 hc$ where ϵ_0 is numerically unity, and represents merely the dimensions of ϵ. Similarly the ratio of the Bohr magneton μ_1 to the Bohr unit of angular momentum $(h/2\pi)$ is strictly $\mu_0 \frac{1}{2}(e/m)$, where μ_0 is numerically unity, and represents merely the dimensions of permeability.

(g) The mole is a (variable) unit of mass, equal to the molecular weight in grams. The gram equivalent is a similar (variable) unit of mass, equal to the atomic or molecular weight in grams, divided by the valence.

The various quantities appearing in Tables a and b have been discussed in the preceding sections. No general explanation will be given of the meaning or use of the quantities appearing in Table c, since any adequate explanation would constitute almost a text-book of modern physics and physical chemistry. In the case of the more specialized constants, no explanation is needed by investigators actually working with each constant, and it is to such persons that the data given here will be most useful.

I wish, in conclusion, to direct attention merely to two constants for which the formula used here differs from that normally given. It is customary to use for the speed of the electron in the normal orbit of hydrogen, as given by Bohr's original theory, a value which refers to the nucleus considered as the center of coordinates. This I call $v_0(=\alpha c)$ in Table c. It would seem more logical to give the speed referred to the center of mass, the quantity denoted v_0' in Table c. There is a similar discrepancy in the case of the radius of this orbit. The electron, according to Bohr, moves about the center of mass in a circle of radius a_0', as it is denoted in Table c. This is not the same as the constant separation of the nucleus and electron, which is here denoted a_0. In the literature these two quantities, a_0, and a_0', are sometimes confused. The expressions for v_0, v_0', a_0, and a_0' given in Table c include also the factor $(1-\alpha^2)^{1/2}$, arising from the variation of mass with speed.

Section O

Conclusions

This final section is devoted to a brief discussion of the status of the general physical constants, as it appears to the writer at the present time (April 1929). As noted in the Introduction, the values of all constants adopted in sections A to M have been based on material available on January 1, 1929. In the present section one can scarcely refrain from mentioning more recent material, especially if it has any direct bearing on the conclusions already reached. Some of this new material has been given in footnotes in the preceding sections. Each section, from A to M, will be considered in order.

(A) It seems unlikely that a more accurate value of the velocity of light, c, will be required for some years to come. In the important relations to be

considered in this concluding section, the probable error in c is far less than that in any of the other constants. Hence, as far as such relations are concerned, the present value of c is entirely satisfactory, and can be considered as more or less permanently established.

(B) The constant of gravitation, G, although of great intrinsic importance, has not been used in later sections of the present paper. The value obtained by Heyl is so much more accurate than any previously available that it also is likely to remain unchanged for some time in the future.

(C) The relation of the liter to the cubic decimeter has not been investigated for many years. The value adopted in section C seems to be universally used, and can be considered as essentially permanent.

(D) The normal mole volume of an ideal gas, v_n, has been actively investigated in recent years. The value adopted is based on 1928 calculations, and it is quite possible that additional work will, at any time, change the most probable value. Although the Baxter and Starkweather investigations, on which the adopted value is founded, are generally considered more trustworthy than any previous work, their results differ somewhat from the previous values and hence a future change in v_n as great as, or even greater than, the adopted probable error would not be surprising.

(E) As has been pointed out in section E, the exact evaluation of the relation between the international and absolute electrical units is a matter of great experimental difficulty, especially since there is no one official standard. Each national laboratory has its own international standards and these may change with time. Progress is being made toward the adoption of the *absolute* electrical units as standards, but for the purpose of correcting work done in terms of the present international units, it is necessary to know the relation between the two systems. The int. ampere differs from the abs. ampere by so small an amount, if at all, that it makes practically no difference whether the ratio q is taken to be unity or not. The int. ohm does, however, definitely differ from the abs. ohm by about one part in 2000 (as shown by the value of p). Because experimental physicists have usually ignored the distinction, many slight changes in published results have been necessary in the later sections of this paper. Due to the variability of the international standards, it is possible that the proper value of p to be used in each specific correction may differ more or less from the adopted value, but the entire possible variation is not more than 2 or 3 parts in 10^5, and this is small compared to the 50 parts in 10^5 by which p differs from unity. The corrections made in this paper are therefore believed to be essentially correct and are not likely to be contradicted by future work.

(F) Of the atomic weights considered in section F, only that of calcium is at all uncertain. This atomic weight has been used in connection with the grating space of calcite and also in the evaluation of the electronic charge e by x-ray measurements. The adopted probable error (0.005) represents only 5 parts in 10^5 in the molecular weight of calcite and thus would have remained negligibly small even if a value several times as large had been adopted.

In this same section evidence was presented to show the agreement of Aston's atomic weights with the chemical values, to at least one part in 10,000, thus indicating that the new oxygen isotope of mass 18 had an abundance of less than one part in 1000 (more accurately, one in 1300) compared to oxygen 16. I have now learned from H. D. Babcock that experiments he has performed on absorption coefficients in the atmospheric bands indicate that O_{18} has an abundance* of only one part in 1250, the probable error being some 25 percent. Hence Aston's atomic weights should be greater than the chemical values by about one part in 10,000. Babcock's determination of relative abundance, however, involves the *assumption* that the absorption coefficient is the same, per molecule, for each species of molecule ($O_{16}-O_{16}$ and $O_{16}-O_{18}$), and this may not be true.

As indicated in footnote 21a of section F, a very recent accurate determination has been made of the atomic weight of silver, directly in terms of oxygen, with a resulting value and error identical with that adopted. The atomic weight of nitrogen therefore becomes unnecessary, for the purposes of this paper.

(G) The normal atmosphere, A_n, is the most accurately known of all the constants considered here. In one sense it is a conventional constant, rather than a general physical constant, since its value involves the *conventional* constant of normal gravity. In experimental work one must measure the actual value of gravity at the position of the apparatus, and this involves its own error. The value of A_n also involves the density of mercury, and since mercury has several isotopes, there are possibilities of uncertainty here, not apparent on the face of the results. Attention is directed also to the "Note added in proof," at the end of section G.

(H) The absolute temperature of the ice-point has an error which is not at all negligible in certain expressions in which it occurs. In fact this accounts for nearly the entire error in the gas constant R_0. As noted in section H, the work of Henning and Heuse indicates an apparent probable error less than the 0.03°C adopted by them and also by the writer, but their result differs from that of Roebuck by 0.04°C. It is quite possible that further work may slightly modify the adopted value and its probable error. There is, however, no question but that the present value of 273.18° is much more trustworthy than the value of 273.09° used until very recently.

(I) The recent precision measurements of the mechanical and of the electrical equivalent of heat are in beautiful agreement and this agreement forms strong evidence that each is substantially correct. It should be remembered, however, that the writer has recalculated the published values of both investigations, and the very fact that the final result depends so much on the method of computation, lends a certain element of uncertainty to the situation. Regardless of this, the present adopted values of J and J' are certainly much more trustworthy than the values based on earlier work.

* This is a corrected figure in which allowance has been made for the fact that the $O_{16}\cdot O_{18}$ molecule has twice as many possible levels as the $O_{16}\cdot O_{16}$ molecule.

(J) As has been pointed out in section J, the value of the Faraday, F, and its adopted error, depend somewhat on the definition of this constant, and there has been some confusion on this point. There are many possibilities of small systematic error in electrolytic work, and these directly affect the value of the Faraday. The writer is aware, from personal correspondence, that his adopted value of F in abs. coulombs (96489 ± 7) is not approved by certain workers in this field. He believes, however, that no valid arguments have been presented to show that it is not the most probable value according to present evidence, and he feels that the true value is likely to differ from the adopted value by less than 7 coulombs rather than by more.

(K) (L) (M). The constants e, e/m, and h, considered in these three sections, together with the velocity of light, appear in many important derived constants. Due to the fact that h depends for its value on e, this latter quantity appears implicitly, if not explicitly, in every quantum relation. The detailed results considered in section M have shown that the value of h would be known with considerable certainty, were it not for the probable error in e. The most outstanding discrepancy in section M was that between the work of Wagner and of Duane and co-workers, on the measurement of h from the x-ray continuous spectrum. As stated in footnote 16a, of that section, the very recent work of Feder, using this same method, gives a value of h in exact agreement with that adopted, and also explains Wagner's very low value. The writer now feels that the value of h/e listed in Table c can be assumed with some confidence. The real problem concerns the values of e and of e/m.

The need of adopting two different values of e/m, one resulting from quantum theory experiments, and one from deflection experiments, is very annoying, and is fundamentally unsatisfactory. The same situation seems to be arising in regard to e. Millikan's oil-drop value has been accepted for many years, since it was the only one available. The new work on x-rays has now opened up another possibility. The value of Bäcklin, considered in some detail in section K, is one-half percent higher than my adopted value. Within the past few days I have been informed by Professor A. H. Compton of the final result of the work of Dr. J. A. Bearden, to which anonymous reference was made in the last paragraph of section K. Dr. Bearden obtains for the absolute wave-length of the (unresolved) Cu $K\alpha$ line, 1.5439 ± 0.0002A, and for the Cu $K\beta$ line, 1.3940 ± 0.0002A. These results are obtained from many different plates, each containing several orders of spectrum, and with three different gratings. The first is 0.345 percent higher than Siegbahn's value, the second 0.336 percent higher. The *relative* wavelengths are thus in entire agreement with Siegbahn, but the *absolute* wavelengths lead to an average value for calcite of $d'_{20} = 3.0398$A, and from Eq. (10), section K, $e = 4.825 \times 10^{-10}$ abs. *es* units. This is 1.15 percent above my adopted value of e. Before expressing any opinion as to this result, it is desirable to consider the various relations that have been suggested as holding between the constants now under consideration.

The most famous of these relations is that connecting e, e/m, h, and c, in Bohr's formula for the Rydberg constant, as stated in Eq. (1) of section M. This relation was used to evaluate h, and the value so obtained (6.54713) is identical to the first four digits with that finally adopted. Hence, in Table *b*, the indirectly calculated value of e/m is also practically identical with that adopted. In other words, the *adopted* values of e, e/m, h and c form a self-consistent system, as judged by the Bohr formula for R_∞. The writer[1] attempted in 1926, to obtain a similarly consistent set of constants, and at that time, in order to accomplish this, altered arbitrarily the value of e by a small amount. This however, was *not* a legitimate procedure, since the necessarily resulting change in h was not made. Fortunately the difficulty has now vanished, assuming the adopted values of the present paper to be correct.

Lewis and Adams,[2] from their theory of ultimate rational units, have obtained, with the aid of Planck's black body radiation law, the following relation between h, c and e,

$$\frac{hc}{2\pi e^2}=8\pi\left(\frac{8\pi^5}{15}\right)^{1/3}. \tag{1}$$

The right side of this equation equals 137.348, while the left side, with the constants here adopted equals $137.29_4\pm 0.11$. The left side, it should be noted, equals the reciprocal of the fine structure constant α, and the value just quoted is taken directly from Table *c*. The numerical agreement of the two sides of Eq. (1) is very striking. This equation was used by the writer,[3] in 1919, as one method for calculating h. The present agreement shows that this method now yields a value of h almost identical (6.5496, to be precise) with that adopted.

The fact that α is a dimensionless constant involving fundamental general constants has often been noted, although it should be remembered, as stated in section N, that to make α strictly dimensionless, we must include with the factor hc the unknown dimensions of specific inductive capacity. H. S. Allen[4] has considered quite fully the various proposed relations of α to other quantities.

Recently Perles[5] has pointed out that the ratio of the mass of the proton to that of the electron (M_p/m_0) is another dimensionless constant which should have some significance, and has found that

$$\frac{hc}{e^2}\left(=\frac{2\pi}{\alpha}\right)=\frac{M_p/m_0}{\pi-1}. \tag{2}$$

With the values adopted here, the left side of Eq (2) equals 862.64 ± 0.68, while the right side equals 858.36 ± 0.49 or 862.26 ± 0.99 depending on whe-

[1] R. T. Birge, Science **64,** 180 (1926). See footnote 8, section M.
[2] G. N. Lewis and E. Q. Adams, Phys. Rev. (2) **3,** 92 (1914).
[3] R. T Birge, Phys. Rev. **14,** 361 (1919).
[4] H. S. Allen "The Quantum and its Interpretation" (1928), pp. 171–183.
[5] J. Perles, Naturwiss. **16,** 1094 (1928).

ther one uses the spectroscopic or the deflection value of e/m. The agreement is evidently very good for the deflection value, but poor for the spectroscopic value. Whether it has any theoretical significance remains to be seen.

A far more important piece of work is Eddington's recent calculation[6] that $1/\alpha = 136$ exactly, from considerations of the new quantum mechanics. This is to be contrasted with our value of 137.29 ± 0.11. The writer[7] has already pointed out briefly this conflict between Eddington's value of α and that deduced from the assumed values of h, c and e. It is desirable now to consider the question more specifically. The equation for $1/\alpha$ can be written

$$\frac{1}{\alpha} = \left(\frac{h}{e}\right)\left(\frac{c}{2\pi e}\right). \tag{3}$$

As has been noted more than once, h/e is known much more accurately than e. Assuming $1/\alpha = 136$, and assuming the adopted value of h/e, we obtain $e = 4.815 \times 10^{-10} es$ units. This is equivalent to assuming that h is to be varied directly as e, and , with the exception of the Rydberg constant method, most of the other methods *do* involve such a direct proportionality. On the other hand, if one assumes my average variation of h and e (the 1.26 power), then $1/\alpha = 136$ requires that $e = 4.8308 \times 10^{-10}$ (and that $h = 6.6520 \times 10^{-27}$). The value of e obtained by Bearden, from x-rays, lies, strangely enough, between the two values just given, and so agrees as well as can be expected with Eddington's value of $1/\alpha$. Hence, so far as Eq. (3) is concerned, we must conclude that if the oil-drop value of e is correct within the assumed probable error, or even within several times that amount, Eddington's value of α, and Bearden's value of e, cannot possibly be correct. On the other hand, if Bearden's value of e, is correct, then Eddington's value of α is also correct, and the oil-drop value of e is wrong.

Direct evidence on these two alternatives is furnished by the Rydberg constant. Eq. (1) of section *M* can be written as

$$R_\infty = \frac{\alpha^2}{2}\left(\frac{e}{h}\right)\frac{1}{(e/m)}. \tag{4}$$

Using our adopted value of e/h and its adopted error, together with Eddington's value of α, one obtains $e/m = (1.7948 \pm 0.0007) \times 10^7$ abs. *em* units, as contrasted with our adopted value of (1.761 ± 0.001), obtained from similar quantum relations. It has already been pointed out that the adopted value of e/m, together with the adopted values of α and e/h, exactly satisfies Eq. (4). The figures just quoted indicate the impossiblity of satisfying Eq. (4)with Eddington's values of α. As already suggested, the discrepancy of 0.5 percent between the spectroscopic and deflection values of e/m may be due to some general error in quantum theory formulas. The above discrepancy of nearly two percent is evidently far more serious In his derivation of the value of $1/\alpha$, Eddington assumes the truth of Eq. (4). Hence the discrepancy

[6] A. S. Eddington, Proc. Roy. Soc. **A122,** 358 (1929).
[7] R. T. Birge, Nature **123,** 318 (1929)

cannot be attributed to a possible theoretical incorrectness of this formula In fact, if Eddington's value of $1/\alpha$ (and by implication Bearden's value of e) is right, one can only conclude that all measurements of e/m, by any method, are too low by 1.5 to 2 percent, and in addition that our adopted value of e is too low by more than one percent. It seems to the writer—an opinion apparently shared by others—that it is more probable that the x-ray values of e are incorrect due to a difference in the structure of a crystal (density, etc.) near the surface from that in the interior.

It must be admitted that this is a very unsatisfactory way to leave the situation in the case of the most important constants known to science. It is to be hoped that before another year has passed, some if not all of these difficulties will have been removed.

THE CORPUSCULAR PROPERTIES OF LIGHT

By Arthur H. Compton
University of Chicago

THE development of the modern conception of light quanta, or photons, began with Planck's ideas concerning heat radiation. Newton indeed had defended the hypothesis of light corpuscles, but the facts which he cited to support this view were later reconciled by Fresnel with the wave theory of light. It was not until new problems were studied, such as the intensity heat radiation and the electrical effects of light, that any real need arose for corpuscles as alternative or supplementary to the wave theory of light.

Planck's Theory of Heat Radiation

Planck was confronted[1] with the fact that the only theory of emission of radiation from hot bodies to which the classical mechanics and electrodynamics would lead, predicted rays much more intense than are actually observed, and of the wrong color. It is a matter of common experience that as a body gradually becomes hotter it first glows a dull red, then orange, and bright gold and finally white. According to the formula developed from the usual kinetic theory, however, the light emitted should always be of the same blue color, differing only in intensity as the temperature changes. Such a conclusion followed necessarily from the fact that all oscillators in thermal equilibrium with each other should have on the average the same kinetic energy, whatever their natural frequency of oscillation. But the oscillators of higher frequency will be subject to greater acceleration if their kinetic energy is the same, and hence, according to electrodynamical principles, should radiate more energy than those of lower frequency. Thus at all temperatures the theory predicted that the high frequency radiation should be more intense than the low frequency radiation.

Planck saw a possible way of escape from this difficulty if he were to suppose that at low temperatures only the oscillators of low frequencies could emit radiation, whereas at high temperatures those of higher frequencies could also radiate. In order to accomplish this result he introduced the simple assumption that the oscillators in the hot body can emit radiation only in units, or quanta, whose energy is proportional to the frequency of the radiation, i.e.,

$$E = h\nu, \tag{1}$$

where h is the constant of proportionality between the frequency and the the energy E of the unit. With this limitation it is possible for only those

[1] Planck, Verh. d. Deut. Phys. Ges. **2**, 237 (1900). A complete account of Planck's studies of this problem is given in his "Wärmestrahlung" (1915), published by Blackiston's in English translation by Masius.

oscillators which have energy greater than $h\nu$ to emit a unit of radiation. Thus at low temperatures, where the average energy of the oscillators is low, only low frequency rays can be emitted. At higher temperatures the higher frequency oscillators will have enough energy to emit their larger units of radiant energy, and so as the temperature rises the center of gravity of the radiation will shift to higher and higher frequencies. Thus with one bold assumption regarding the unitary nature of the emitted light, Planck was able to arrive at a reasonable explanation of the hitherto insoluble problem of the color of the light emitted by hot bodies.

It would take us too far afield to describe how Planck developed this idea of energy quanta to account quantitatively for the intensity as well as the spectral energy distribution of heat rays. In his hands and those of others the theory has assumed a variety of forms, but it has always retained the essential feature that the rays from the hot body must be emitted in units whose energy is proportional to the frequency. The introduction of this idea has marked the opening of an important epoch in the development of theoretical physics.

Einstein and the Photoelectric Effect

The units of radiant energy introduced by Planck were not corpuscular. He supposed that the radiation from an oscillator, though having a definite amount of energy, would spread itself through all space after the manner of a spherical electromagnetic wave. It remained for Einstein[2] to introduce the conception of a corpuscular unit of radiation, or photon, in his effort to account for the photoelectric effect.

When Einstein approached this problem it was recognized that the speed with which photoelectrons are ejected from a surface increases with increasing frequency of the light, and it was generally supposed that the number of photoelectrons emitted was proportional to the intensity of the light striking the photoelectric surface. He saw that this proportionality would follow from the assumption that the light which excited the photoelectrons occurs as a stream of particles, each of which would spend its energy in ejecting an electron fron an atom of the photoelectric material. If each of these particles had energy $h\nu$, as might be inferred from Planck's theory of heat radiation, this picture of the process would account also for the increase of speed with higher frequencies. If a certain amount of work w_0 is required to remove the electron from the atom, Einstein supposed that all the rest of the photon's energy is spent in giving kinetic energy to the electron, thus deriving his famous photoelectric equation,

$$E_{kin} = h\nu - w_0. \tag{2}$$

It was years before this theory received an adequate test. Experiments by Ladenburg[3] favored the view that the velocity rather than the kinetic energy

[2] A. Einstein, Ann. d. Physik **17**, 145 (1905).
[3] E. Ladenburg, Phys. Zeits. **8**, 590 (1907).

was proportional to the frequency of the incident light, and different results were obtained with different metals. Richardson and Compton[4] and independently Hughes[5] showed that the differences found for different metals were due to their different contact potentials, and to the fact that the value of w_0 is different from metal to metal. They were indeed able to show that Einstein's equation was of the right form, and that the constant of proportionality h is approximately the same as Planck's constant. A few years later Millikan,[6] using greater care in securing strictly monochromatic light, was able by means of Einstein's equation to secure from photoelectric measurements one of the best experimental determinations that we have of Planck's constant.

The photoelectric effect is especially prominent with x-rays, for these rays eject photoelectrons from all kinds of substances. The velocities of these x-ray photoelectrons have been measured by means of their curvature in a magnetic field, using the so-called magnetic spectrograph. M. de Broglie[7] showed that even for these very high frequencies Einstein's equation holds, if by w_0 we now mean the work required to remove the electron from the o level of the atom. In fact Robinson[8] has applied this equation to his measurements of the speed of x-ray photoelectrons from various substances as a powerful method of studying the energy levels of the different atoms. In a similar way, Ellis,[9] Meitner,[10] Thibaud[11] and others have used equation (2) as a means of determining γ-ray frequencies from the speed of the secondary β-rays. Recent measurements of these frequencies by crystal methods[12] show that even for these exceedingly great energies Einstein's law holds. Over a range of kinetic energies corresponding to a drop through potential differences from 1 volt to 2 million volts Einstein's photoelectric equation has thus been verified to within an experimental error of 1 percent. It is thus one of the most adequately tested laws in the realm of physics.

Before these photoelectric experiments had been carried to a successful conclusion, Duane and Hunt[13] observed a closely related phenomenon which is frequently called the inverse photoelectric effect. They found that when an x-ray tube is operated at a constant potential, there is a definite lower limit to the wave-length of the x-rays from the tube, and that this limiting wave-length is inversely proportional to the voltage. This result may be written in the form,

$$Ve = hc/\lambda_{min} = h\nu_{max}, \tag{3}$$

[4] O. W. Richardson and K. T. Compton, Phil. Mag. **24,** 575 (1912).
[5] A. L. Hughes, Phil. Trans. Roy. Soc. **A212,** 205 (1912).
[6] R. A. Millikan, Phys. Rev. **7,** 18 and 355 (1916).
[7] M. de Broglie, J. de Phys. et Radium **2,** 265 (1921).
[8] H. R. Robinson, Phil. Mag. **50,** 241 (1925).
[9] C. D. Ellis, Proc. Roy. Soc. **A100,** 1 (1922); Proc. Camb. Phil. Soc. **22,** 374 (1924), et al.
[10] L. Meitner, Zeits. f. Physik **11,** 35 (1922), et al.
[11] J. Thibaud, Comptes rendus **178,** (1924), et al.
[12] L. T. Steadman, Phys. Rev. **33,** 1069 (1929).
[13] W. Duane and F. L. Hunt, Phys. Rev. **6,** 166 (1915).

where V is the applied potential, and the other letters have their usual significance. Duane and Hunt's quantitative measurements, confirmed and extended by a number of other investigators[14] have shown that the factor of proportionality in this equation is the same quantity h as that which appears in Planck's theory, equation (1). In fact the measurement of this limiting x-ray wave-length is perhaps our best direct method of determining Planck's constant.

The significance of this work will perhaps be more obvious if we imagine the following experiment: Let two x-ray tubes, A and B, be placed side by side. Tube A is operated at a constant potential of say 100,000 volts. A cathode electron with a kinetic energy Ve strikes the target of tube A and gives raise to an x-ray of frequency $\nu = Ve/h$. This ray strikes the target of tube B and there ejects a photoelectron whose kinetic energy according to equation(2) is $Ve - w_0$. This means that all of the energy of the cathode electron in tube A has been transmitted to the photoelectron ejected from the target of tube B. How is it possible for such a complete transfer of energy to be effected?

A precisely similar difficulty arises in connection with Bohr's picture of radiation and absorption by atoms, which was developed[15] while these studies of the photoelectric effect were going on. According to this picture, radiation is emitted by an atom only when it changes from one state to another having less energy, in which case the frequency is given by the expression,

$$h\nu = \delta E, \tag{4}$$

where δE is the loss in energy by the atom. When an atom absorbs energy, it changes from one state to another of higher energy and the frequency of the absorbed radiation is again given by equation (4), where δE now means the increase in the energy of the atom. Thus we see again that if one system suddenly radiated an amount of energy δE, another atomic system, which may be as far away as the earth is from a distant star, may suddenly have its energy increased by the same amount when the radiation reaches it.

The impossibility that an electromagnetic wave whose energy spreads in all directions should effect such a sudden and complete transfer of energy is obvious. It is equally clear that Einstein's photon conception affords a simple and adequate method of making the transfer. There have not been lacking, however, attempts to explain these phenomena without resorting to assumptions departing so completely from the electromagnetic waves of Maxwell.

One such attempt is the accumulation hypothesis, according to which the light energy is gradually accumulated by the atom, and the photoelectron is finally ejected when the accumulated energy exceeds a certain critical value. This process requires the existence of stored energy of all possible amounts within the atom, since the kinetic energy of the ejected photo-

[14] *E. g.*, Duane, Palmer and Chi-Sun-Yeh, J. Opt. Soc. Am. **5**, 376 (1921); E. Wagner, J. d. Rad. Elek. **16**, 212 (1919).

[15] N. Bohr, Phil. Mag. **26**, 1,476 and 857 (1913).

electron may have any value, depending upon the frequency of the radiation which traverses its parent atom. Furthermore this energy must remain stored for indefinitely long periods of time, for otherwise emission of photoelectrons would not occur at once upon exposure to the light—time would be required for the atom to accumulate sufficient energy. We are thus led to imagine an atom which may possess any energy whatever, and whose energy may gradually increase as radiation is absorbed. Such a picture is wholly inconsistent with Bohr's idea of an atom with definite stationary states and which changes only suddenly from one such state to another. It is true that recent developments in quantum mechanics have led us to revise considerably Bohr's conception of electron orbits; but this hypothesis of stationary states seems more firmly established than ever, and continues to be the fundamental principle of spectral analysis. We thus find it difficult to consider seriously an accumulation hypothesis which would mean atoms having all possible amounts of energy.

There is another apparently fatal difficulty with this explanation of the photoelectric effect, in that it fails to account for the direction in which the photoelectrons are emitted. Experiments by Wilson[16] Bubb[17] and others[18] have shown that the most probable direction in which a photoelectron is ejected from an atom by x-rays is nearly that of the electric vector of the incident wave, but with an appreciable average forward component to its motion. This forward component is about equal to the momentum $h\nu/c$ of the incident photon, as would be expected if the electron suddenly absorbs a photon of energy $h\nu$ and escapes before any appreciable impulse has been transferred to the atom.[19] On the other hand, if the energy is gradually accumulated by the electron, the forward impulse received from the radiation would be transferred to the whole atom, and no reason appears for the strong forward component to the photoelectron's motion. Thus the accumulation hypothesis does not seem to be tenable.

If the atom cannot gradually accumulate energy, since a spherical electromagnetic wave cannot give up its whole energy to a single atom, the occurrence of photoelectrons with the energy $h\nu$ means that we must either give up our old view that light comes in spherical waves or abandon the doctrine of the conservation of energy. Bohr, Kramers and Slater[20] at one time preferred to assume that energy is not conserved when an individual photoelectron is produced. They supposed that on the average the energy appearing in the photoelectrons is equal to that absorbed from the radiation, but under

[16] C. T. R. Wilson, Proc. Roy. Soc. **A104,** 1 (1923).

[17] F. W. Bubb, Phys. Rev. **23,** 137 (1924).

[18] *E.g.*, Auger, Comptes rendus **178,** 1535 (1924); D. H. Loughridge, Phys. Rev. **26,** 697 (1925); F. Kirchner, Zeits. f. Physik **27,** 285 (1926); E. J. Williams, Nature **121,** 134 (1928).

[19] The average forward component is found in certain recent experiments to approach a value $9/5 \times h/\lambda$, where h/λ is the momentum of the photon. This value has been derived on the basis of wave mechanics (cf. A. Sommerfeld "Atombau Ergänzungsband" 1929 p. 222, and E. J. Williams, Nature **123,** 565, (1929).

[20] N. Bohr, H. A. Kramers and J. C. Slater, Phil. Mag. **47,** 785 (1924).

the stimulus of the incident waves any particular electron might suddenly escape at high speed without any corresponding loss in energy by the remainder of the system. That is, the conservation of energy, and similarly the conservation of momentum, would become statistical laws. The authors of this theory assume that, though the rays are propagated as spherical waves the motion of the photoelectrons would be the same as if they were ejected by photons. It has thus been difficult to devise a photoelectric experiment which would distinguish between this "virtual radiation"hypothesis and that of photons. The degree of success that has attended the application of the photon hypothesis to the motion of photoelectrons has however come directly from the application of the conservation principles to the individual action of a photon on an electron. The power of these principles as applied to this case is surprising if the assumption is correct that they are only statistically valid.

Quantum Phenomena Associated with the Scattering of X-Rays

We have seen that Einstein's hypothesis of corpuscular units of radiant energy gives a satisfactory account of the photoelectric effect. As Jeans has significantly remarked, however, Einstein invented the photon hypothesis just to account for this one effect, and it is not surprising that it should account for it well. In order to carry any great weight the hypothesis should also be found applicable to some phenomena of widely different character. Just such phenomena have recently been found associated with the scattering of x-rays—the change in wave-length of the scattered rays, and the recoil electrons associated with them.

The earliest experiments on secondary x-rays and γ-rays showed a difference in the penetrating power of the primary and the secondary rays. Barkla[21] and his collaborators showed that the secondary rays from the heavy elements consisted largely of fluorescence radiations characteristic of the radiator, and that it was the presence of these softer rays which was chiefly responsible for the great absorption of the secondary rays. When later experiments showed a measureable difference in penetration even for light elements such as carbon, from which no fluorescence radiation appears, it was natural to ascribe this difference to a new type of fluorescence radiation, similar to the K and L types, but of shorter wave-length.[22] Careful absorption measurements[23] failed however, to reveal any critical absorption limit for these assumed "J" radiations similar to those corresponding to the fluorescence K and L radiations. Moreover, direct spectroscopic observations[24] failed to reveal the existence of any spectrum lines under conditions with which the supposed J rays should appear. It thus became evident that the softening of the secondary x-rays from the lighter elements was due to a different

[21] C. G. Barkla and C. A. Sadler, Phil. Mag. **16,** 550 (1908).

[22] C. G. Barkla and Miss White, Phil. Mag. **34,** 270 (1917); J. Laub, Ann. d. Physik **46,** 785 (1915); J. A. Crowther, Phil. Mag. **42,** 719 (1921).

[23] Richtmyer and Grant, Phys. Rev. **15,** 547 (1920).

[24] Duane and Shimizu, Phys. Rev. **13,** 288 (1919); **14,** 389 (1919).

kind of process from the softening of the secondary rays from heavy elements where fluorescence x-rays are present.

According to the usual electron theory of x-ray scattering, the primary waves set the electrons into forced oscillation, and they, because of their accelerations, reradiate x-rays in all directions. It is thus obvious that the scattered rays will be of the same frequency as the primary rays which set the electrons in motion. A series of skillfully devised absorption experiments, performed by J. A. Gray,[25] showed however that both in the case of x-rays and γ-rays an increase in wave-length accompanies the scattering of the rays by light elements. When spectroscopic studies were made[26] they likewise revvealed lines in the spectrum of the scattered rays corresponding to those in the primary beam, but with each line displaced slightly toward the longer wave-lengths. These spectra had the advantage over the absorption measurements of affording a quantitative determination of the change in wavelength, which gave a basis for its theoretical interpretation.

The photon conception gives a simple interpretation of this phenomenon. If we suppose that each x-ray photon is deflected by a single electron the electron will recoil from the impact. That is, part of the photon's energy is spent in setting the electron in motion, so the photon has less energy after deflection than before. The problem is very similar to that of the elastic collision of a light ball with a heavy one. If we assume that the energy and momentum are conserved in the process we can calculate the loss in energy and hence the increase in wave-length of a photon which is scattered at an angle ϕ with the primary ray. We thus find[27] for the increase in wavelength,

$$\delta\lambda = \frac{h}{mc}(1-\cos\phi), \tag{5}$$

where h is again Planck's constant, m is the mass of the electron and c is the velocity of light. The electron at the same time recoils from the photon at an angle θ given by,

$$\cot\theta = -(1+\alpha)\tan\tfrac{1}{2}\phi, \tag{6}$$

where $\alpha = h/mc\lambda$, and the kinetic energy of the recoiling electron is

$$E_{kin} = h\nu\frac{2\alpha\cos^2\theta}{(1+\alpha)^2-\alpha^2\cos^2\theta}. \tag{7}$$

Until very recently the experiments showed just two lines in the spectrum of the scattered rays corresponding to each line of the primary ray. Of these lines one, the "unmodified line," is of very nearly the same wavelength as the primary ray, whereas the second, or "modified line," though apparently somewhat broadened, has its center of gravity shifted by ap-

[25] J. A. Gray, Phil. Mag. **26,** 611 (1913); J. Frank. Inst. Nov. 1920, p. 643.
[26] A. H. Compton, Bull. Nat. Res. Coun. No. 20 (1922); Phys. Rev. **22,** 409 (1923).
[27] A. H. Compton, Phys. Rev. **22,** 483 (1923); P. Debye, Phys. Zeits. **24,** 161 (1923).

proximately the amount predicted by equation (5). According to experiments by Kallman and Mark[28] and by Sharp[29] the agreement between the theoretical and the observed shift is precise to within a small fraction of 1 percent.

Within the last year Davis and Mitchell,[30] using their high resolving power double crystal spectrometer, have found that the "unmodified line" has a complex structure, with one line the same wave-length as the primary rays and with a group of other lines each of whose frequencies differs from that of the primary beam by approximately the limiting frequency of some energy level in the normal atom. Thus there is a line whose frequency is given approximately by the relation,

$$h\nu'' = h\nu - h\nu_k \tag{8}$$

where ν is the primary frequency and $h\nu_K$ is the energy of the K energy level Such lines may be accounted for by supposing that the incident photon spends enough energy upon the atom to release a K electron (or to transfer it to an outside orbit) and then escapes from the atom with its remaining energy. The process is thus analogous to the photoelectric effect, where, however, the photon instead of the electron escapes with the energy remaining after the electron has been removed from its original orbit. These lines seem to be the x-ray analogues of the Raman lines, which had been discovered[31] a few months earlier in the visible spectrum.

Very recent experiments by Davis and Purks[32] have shown a similar fine structure in the modified line. Such a structure is consistent with the photon conception of the scattering process, and had indeed been predicted on this basis by the writer[33] using certain assumtions regarding the action of the photon on bound electrons. The experiments of Sharp[29] and DuMond,[34] however, seem to indicate a broad and almost structureless modified line, which would seem to be in disagreement with the results of Davis and Purks. The spectra obtained by the latter investigators also show a change in wave-length which is almost 10 percent less than that predicted by equation (5), a result difficult to reconcile with the observations of other recent experiments. It is important to settle these differences in the detailed experimental results, because they are of significance regarding the manner in which a photon acts upon an electron bound within an atom. There is in these experiments, however, no evidence of a disagreement with the basic corpuscular theory from which equation (5) was derived.

[28] H. Kallman and H. Mark, Naturwiss. **13,** 297 (1925).
[29] H. M. Sharp, Phys. Rev. **26,** 691 (1925).
[30] B. Davis and D. P. Mitchell, Phys. Rev. **32,** 331 (1928).
[31] C. V. Raman, Indian J. Phys. **2,** 387 (1928).
[32] B. Davis and H. Purks, Phys. Rev. **34,** 1 (1929).
[33] A. H. Compton, Phys. Rev. **24,** 168 (1924).
[34] J. W. M. DuMond, Phys. Rev. **33,** 643 (1929).

Recoil Electrons

From the close agreement between the theoretical and the observed wave-lengths of the scattered rays, the recoil electrons predicted by the photon theory of scattering were looked for with some confidence. When this theory was proposed, there was no direct evidence for the existence of such electrons, though indirect evidence suggested[35] that the secondary β-rays ejected from matter by hard γ-rays are mostly of this type. Within a few months of their prediction, however, C. T. R. Wilson[36] and W. Bothe[37] independently announced their discovery. The recoil electrons show as short tracks in the cloud expansion photographs, pointed in the direction of the primary beam, mixed among the much longer tracks due to the photo-electrons ejected by the x-rays.

Perhaps the most convincing reason for associating these short tracks with the scattered x-rays comes from a study of their number. Each photo-electron in a cloud photograph represents a quantum of truly absorbed x-ray energy. If the short tracks are due to recoil electrons, each one should represent the scattering of a photon. Thus the ratio N_r/N_p of the number of short tracks to the number of long tracks should be the same as the ratio σ/τ of the scattered to the truly absorbed energy when the x-rays pass through air. The latter ratio is known from absorption measurements, and the former ratio can be determined by counting the tracks on the photographs. The satisfactory agreement between the two ratios for x-rays of different wave-lengths means that on the average there is about one quantum of energy scattered for each short track that is produced.

This result is in itself contrary to the predictions of the classical wave theory, since on this basis all the energy spent on a free electron (except the insignificant effect of radiation pressure) should reappear as scattered x-rays. In these experiments on the contrary, 5 or 10 percent as much energy appears in the motion of the recoil electrons as in the scattered x-rays.

That these short tracks correspond to the recoil electrons predicted by the photon theory of scattering becomes clear from a study of their energies. The energy of an electron which produces a track in an expansion chamber can be calculated from the range of the track. The ranges of the tracks which start in different directions have been studied,[38] using primary x-rays of different wave-lengths, with the result that equation (7) has been satisfactorily verified. A more accurate check on these recoil electron energies has recently been made by Bless,[39] using a magnetic spectrometer, and with results wholly consistent with the theory.

In view of the fact that electrons of the recoil type were unknown when the photon theory of scattering was presented, their existence, and the close

[35] A. H. Compton, Bull. Nat. Res. Coun. No. **20**, p. 27 (1922).

[36] C. T. R. Wilson, Proc. Roy. Soc. **A104**, 1 (1923).

[37] W. Bothe, Zeits. f. Physik **16**, 319 (1923); **20**, 237 (1923).

[38] A. H. Compton and A. W. Simon, Phys. Rev. **25**, 306 (1925); J. M. Nuttall and E. J. Williams, Manchester Memoirs, **70**, 1 (1926).

[39] A. A. Bless, Phys. Rev. **30**, 871 (1927).

agreement with the predictions as to their number, direction and velocity, supply strong evidence in favor of the photon hypothesis.

Interpretation of these Experiments

It is impossible to account for scattered rays of altered frequency, and for the existence of the recoil electrons, if we assume that x-rays consist of electromagnetic waves in the ordinary sense. Yet some progress has been made on the basis of semi-classical theories. It is an interesting fact that the wave-length of the scattered ray according to equation (5) varies with the angle just as one would expect from a Doppler effect if the rays are scattered from an electron moving in the direction of the primary beam. Moreover, the velocity that must be assigned to the electron in order to give the proper magnitude to the change of wave-length is that which the electron would acquire by radiation pressure if it should absorb a quantum of the incident rays. Several writers[40] have therefore assumed that an electron takes from the incident beam a whole quantum of the incident radiation, and then emits this energy as a spherical wave while moving forward with high velocity. There is, however, the difficulty that this theory predicts recoil electrons all moving in the same direction and with the same velocity. The experiments show, on the other hand, a variety of directions and velocities, with the velocity and direction correlated as demanded by the photon hypothesis. Moreover, the maximum range of the recoil electrons, though in agreement with the predictions of the photon theory, is found[38] to be some four times as great as that predicted by this semi-classical theory.

There is nothing in these experiments, however, which is inconsistent with the idea of virtual oscillators continually scattering virtual radiation. In order to account for the change of wave-length on this view, Bohr, Kramers and Slater assumed[20] that the virtual oscillators scatter as if moving in the direction of the primary beam, accounting for the change in wave-length as a Doppler effect. They then suppose that occasionally an electron, excited by the primary virtual rays, will suddenly move forward as if it had received the momentum of a photon. Though we have seen that the electrons move in a variety of different directions, the theory could easily be extended to include the type of motion that is actually observed. It is difficult, however, to see how such a theory could by itself predict the change in wave-length and the motion of the recoil electrons.

We may conclude that the photon theory predicts quantitatively and in detail the change of wave-length of the scattered x-rays and the characteristics of the recoil electrons. The virtual radiation theory is probably not inconsistent with these experiments, but is incapable of predicting the results. The classical theory is altogether helpless to deal with these phenomena.

[40] C. R. Bauer, Comptes rendus **177,** 1211 (1923); C. T. R. Wilson, Proc. Roy. Soc. **A104,** 1 (1923); K. Forsterling, Phys. Zeits. **25,** 313 (1924); O. Halpern, Zeits. f. Physik **30,** 153 (1924).

Experiments with Individual Radiation Quanta

We have seen that while these experiments on the photoelectric effect and on the scattering of x-rays give results which cannot be reconciled with the classical picture of electromagnetic waves, they do not suffice to distinguish between the photon theory and the theory of virtual radiation. The latter theory succeeded in avoiding the difficulties of the classical theory by considering the conservation of energy and momentum as only statistically valid. If experiments can be performed on the interaction of individual photons and electrons, it should be possible to make a direct test of the conservation principles, and to distinguish between the virtual radiation hypothesis and that of photons. Three important experiments of this type have been reported.

(1) *Test for Coincidences with Fluorescence X-Rays.* Bothe has performed an experiment[41] in which fluorescence rays from a thin copper foil are excited by a beam of incident x-rays. Two point counters of the type developed by Geiger are mounted, one on either side of the foil, in each of which an average of 1 photoelectron is recorded for about 20 quanta radiated by the foil. If we assume that the fluorescence radiation is emitted in quanta of energy, but proceed in spherical waves in all directions, there should thus be about 1 chance in 20 that the recording of a photoelectron in one chamber should be simultaneous with the recording of a photoelectron in the other. The experiments showed no coincidences other than those which were explicable by such sources as high speed beta particles which traverse both counting chambers.

This result is in accord with the photon hypothesis. For if a photon of fluorescence radiation produces a β-ray in one counting chamber it cannot traverse the second chamber. Coincidences should therefore not occur.

According to the virtual radiation hypothesis, however, coincidences should have been observed. For on this view the fluorescence K radiation is emitted by virtual oscillators associated with atoms in which there is a vacancy in the K shell. That is, the copper foil can emit fluorescence K radiation only during the short interval of time following the expulsion of a photoelectron from the K shell, until the shell is again occupied by another electron. This time interval is so short (less than 10^{-15} seconds) as to be sensibly instantaneous on the scale of Bothe's experiments. Since on this view the virtual radiation is emitted in spherical waves, the counting chambers on both sides of the foil should be simultaneously affected, and coincident pulses in the two chambers should frequently occur. The results of the experiment are thus contrary to the predictions of the virtual radiation hypothesis.

(2) *Coincidences of Scattered X-Rays and Recoil Electrons.* We have seen according to Bohr, Kramers and Slater's theory, virtual radiation is being continually scattered by matter traversed by x-rays, but only occasionally is a recoil electron emitted. This is in sharp contrast with the photon theory, according to which a recoil electron appears every time a photon is scattered.

[41] W. Bothe, Zeits. f. Physik **37**, 547 (1926).

A crucial test between these two points of view is afforded by an experiment devised and brilliantly performed by Bothe and Geiger.[42] X-rays were passed through hydrogen gas, and the resulting recoil electrons and scattered rays were detected by means of two point counters on opposite sides of the column of gas. The chamber for counting recoil electrons was left open, but a thin sheet of platinum prevented the recoil electrons from entering the chamber for counting the scattered rays. The impulses from the counting chambers were recorded on a moving photographic film.

In observations over a period of five hours 66 coincidences between the impulses due to recoil electrons and the scattered rays were observed. Bothe and Geiger estimated that according to the virtual radiation theory the chance was only 1 in 400,000 that so many coincidences should have occurred. This result is therefore in accord with the predictions of the photon theory, but is directly contrary to the statistical view of the scattering process.

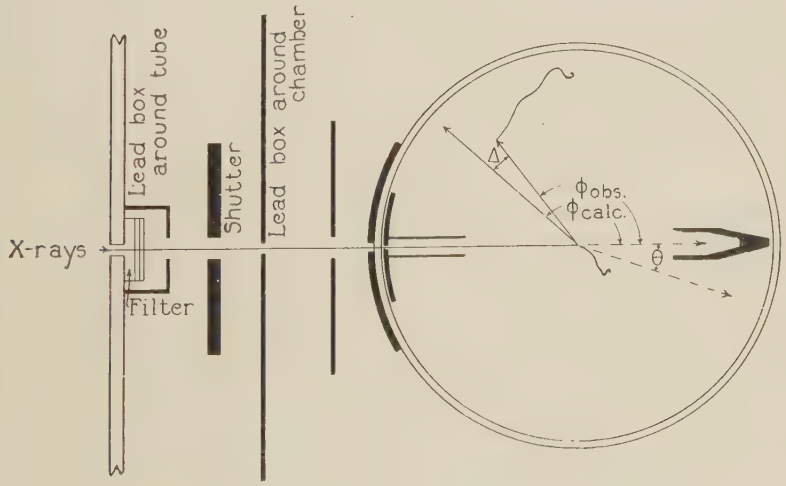


Fig. 1.

(3) *Directional Emission of Scattered X-Rays.* According to the photon theory, we have a definite relation (equation 6) between the angle at which the photon is scattered and the angle at which the recoil electron is ejected. But according to any form of spreading wave theory, including that of Bohr, Kramers and Slater, the scattered rays may produce effects in any direction whatever, and there should be no correlation between the direction in which a recoil electron proceeds and the direction in which the scattered x-ray produces an effect. A test to see whether such a relation exists has been made,[43] using a cloud expansion apparatus, in the manner shown diagrammatically in Fig. 1. Each recoil electron produces a visible track, and oc-

[42] W. Bothe and H. Geiger, Zeits. f. Physik **26,** 44 (1924); **32,** 639 (1925).
[43] A. H. Compton and A. W. Simon, Phys. Rev. **26,** 289 (1925).

casionally a secondary track is produced by the scattered x-ray before it escapes from the chamber. When but one recoil electron appears on the same plate with the track due to the scattered rays, it is possible to tell at once whether the angles satisfy equation (6). If two or three recoil tracks appear, the measurements on each track can be appropriately weighted.

Of 850 plates taken in the final series of readings, 38 show both recoil tracks and secondary β-ray tracks. On 18 of these plates the observed angle ϕ is within 20 degrees of the angle calculated from the measured value of θ, while the other 20 tracks are distributed at random angles. This ratio 18:20 is about that to be expected for the ratio of the rays scattered by the part of the air from which the recoil tracks could be measured to the stray rays from various sources. There is only about 1 chance in 250 that so many secondary β-rays should have appeared at the theoretical angle.

This result means that associated with each recoil electron there is scattered x-ray energy sufficient to produce a β-ray, and proceeding in a direction determined at the moment of ejection of the recoil electron. In other words, the scattered x-rays proceed in directed units of radiant energy.

This result, like those of the previous two experiments, is irreconcilable with the virtual radiation hypothesis of the production of recoil and photoelectrons. On the other hand all of these experiments with individual radiation quanta are in complete accord with the predictions of the photon theory.

The Paradox of Waves and Particles

Experiments on the photoelectric effect and on scattered x-rays, taken together with these experiments on the individual interactions of radiation and electrons, show therefore that radiation is emitted in units, is propagated in definite directions, and is absorbed again in units of undiminished energy. Light thus has all the essential characteristics of particles. It is well known however that light has the characteristics of waves. The phenomena of reflection, refraction, polarization and interference, which occur with light, can leave no reasonable doubt about its wave properties. How can these two apparently conflicting conceptions be reconciled?

Electron Waves. Before attempting to answer this question, let us notice that this dilemma applies not only to radiation, but also to other fundamental fields of physics. When the evidence was growing strong that radiation, which we have always thought of as waves, had the properties of particles, L. de Broglie asked, may it not then be possible that electrons, which we have known as particles, may have the properties of waves? He was able to give a mathematical proof[44] that the dynamics of any particle may be expressed in terms of the propagation of a group of waves. That is, the particle may be replaced by a train of waves—the two, so far as their motion is concerned, may be made mathematically equivalent. The motion of a particle in a straight line is represented by a plane wave. The wave-length is determined by the momentum of the particle. Thus just as the momentum of

[44] L. de Broglie, Phil. Mag. **47,** 446 (1924); Thesis, Paris 1924.

a photon is $h\nu/c = h/\lambda$ so the wave-length of a moving electron is given by $mv = h/\lambda$, or

$$\lambda = h/mv. \tag{9}$$

In C. T. R. Wilson's cloud expansion photographs we have ocular evidence that electrons are very real particles indeed. Nevertheless de Broglie's suggestion that they should act as waves has been subjected to experimental test by Davisson and Germer[45] and later by G. P. Thomson,[46] Rupp,[47] Kikuchi[48] and others.

For our present purpose we may describe Thomson's experiment, which is typical of them all. His experiment is analogous to those in which Debye and Scherrer[49] and Hull[50] secured diffraction patterns of x-rays by passing them through powdered crystals placed some distance in front of a photographic plate. Thomson replaced the x-ray beam with a stream of cathode rays (falling through about 30,000 volts potential difference), and the mass of powdered crystals with a sheet of gold leaf. The resulting photographs showed the same kind of diffraction pattern as that obtained when x-rays pass through gold leaf. Indeed from the size of the diffraction rings the wave-length of the cathode rays could be calculated, and was found to be just that predicted by de Broglie's formula (9). If the diffraction of x-rays by crystals proves that they are waves, this diffraction of cathode rays establishes equally the wave characteristics of electrons.

We are thus faced with the fact that the fundamental things in nature, matter and radiation, present to us a dual aspect. In certain ways they act like particles, in others like waves. The experiments tell us that we must seize both horns of the dilemma.

A Suggested Solution. During the last few years there has gradually developed a solution of this puzzle, which though at first rather difficult to grasp seems to be free from logical contradictions and essentially capable of describing the phenomena which our experiments reveal. A mention of some of the names connected with this development will suggest some of the complexities through which the theory has gradually gone. There are L. de Broglie, Duane, Slater, Schrödinger, Heisenberg, Bohr and Dirac, among others, who have contributed to the growth of this explanation.[51] The point of departure of this theory is de Broglie's proof, mentioned above, that the motion of a particle may be expressed in terms of the propagation of a group of waves. In the case of the photon, this wave may be taken as the ordinary

[45] C. J. Davisson and L. H. Germer, Phys. Rev. **30,** 705 (1927).
[46] G. P. Thomson, Proc. Roy Soc. **A117,** 600 (1928); **A119,** 651 (1928).
[47] E. Rupp, Ann. d. Physik **85,** 981 (1928).
[48] S. Kikuchi, Jap. J. of Phys. **5,** 83 (1928).
[49] Debye and Scherrer, Phys. Zeits. **17,** 277 (1916).
[50] A. W. Hull, Phys. Rev. **10,** 661 (1917).
[51] A review of the development of this theory is given in the report of the fifth Solvay Congress, "Electrons et Photons," Brussels, 1928, written chiefly by W. L. Bragg, A. H. Compton, L. de Brogile, E. Schrödinger, W. Heisenberg and N. Bohr.

electromagnetic wave. The wave corresponding to the moving electron is generally called by the name of its inventor, a de Broglie wave.

Consider, for example, the deflection of a photon by an electron on this basis, that is, the scattering of an x-ray.[52] The incident photon is represented by a train of plane electromagnetic waves. The recoiling electron is likewise represented by a train of plane de Broglie waves propagated in the direction of recoil. These electron waves form a kind of grating by which the incident electromagnetic waves are diffracted. The diffracted waves represent in turn the deflected photon. They are increased in wave-length by the diffraction because the grating is receding, resulting in a Doppler effect.

In this solution of the problem we note that before we could determine the direction in which the x-ray was to be deflected, it was necessary to know the direction of recoil of the electron. In this respect the solution is indeterminate; but its indeterminateness corresponds to an indeterminateness in the experiment itself. There is no way of performing the experiment so as to make the electron recoil in a definite direction as a result of an encounter with a photon. It is a beauty of the theory that it is determinate only where the experiment itself is determinate, and leaves arbitrary those parameters which the experiment is incapable of defining.

It is not usually possible to describe the motion of either a beam of light or a beam of electrons without introducing both the concepts of particles and waves. There are certain localized regions in which at a certain moment energy exists, and this may be taken as a definition of what we mean by a particle. But in predicting where these localized positions are to be at a later instant, a consideration of the propagation of the corresponding waves is usually our most satisfactory mode of attack.

Attention should be called to the fact that the electromagnetic waves and the de Broglie waves are according to this theory waves of probability. Consider as an example the diffraction pattern of a beam of light or of electrons, reflected from a ruled grating, and falling on a photographic plate. In the intense portion of the diffraction pattern there is a high probability that a grain of the photographic plate will be affected. In corpuscular language, there is a high probability that a photon or electron, as the case may be, will strike this portion of the plate. Where the diffraction pattern is of zero intensity, the probability of a particle striking is zero, and the plate is unaffected. Thus there is high probability that a photon will be present where the "intensity" of an electromagnetic wave is great, and a lesser probability where this "intensity" is smaller.

It is a corollary that the energy of the radiation lies in the photons, and not in the waves. For we mean by energy the ability to do work, and we find that when radiation does anything it acts in particles.

In this connection it may be noted that this wave-mechanics theory does not enable us to locate a photon or an electron definitely except at the

[52] E. Schrödinger, Ann. d. Physik **82**, 257 (1927).

instant at which it interacts with another particle. When it activates a grain on a photographic plate, or ionizes an atom which may be observed in a cloud expansion chamber, we can say that the particle was at that point at the instant of the event. But in between such events the particle can not be definitely located. Some positions are more probable than others, in proportion as the corresponding wave is more intense in these positions. But there is no definite position that can be assigned to the particle in between its actions on other particles. Thus it becomes meaningless to attempt to assign any definite path to a particle. It is like assigning a definite path to a ray of light: the more sharply we try to define it by narrow slits the more widely the ray is spread by diffraction.

Perhaps enough has been said to show that by grasping both horns we have found it possible to overcome the dilemma. Though no simple picture has been invented affording a mechanical model of a light ray, by combining the notions of waves and particles a logically consistent theory has been devised which seems essentially capable of accounting for the properties of light as we know them.

Starting with Planck's epoch-making suggestion that radiation is emitted in discrete units proportional to the frequency, we have thus seen how Einstein was led to suggest corpuscular quanta of radiation or photons in order to account for the photoelectric effect, and how recent experiments with x-rays, especially those with individual x-ray quanta, have seemed to establish this corpuscular hypothesis. Yet we have long known that light has the characteristics of waves. For centuries it has been supposed that the two conceptions are contradictory. Goaded on, however, by the obstinate experiments, we seem to have found a way out. We continue to think of light as propagated as electromagnetic waves; yet the energy of the light is concentrated in particles associated with the waves, and whenever the light does something it does it as particles.

STATISTICAL THEORIES OF MATTER, RADIATION AND ELECTRICITY

By Karl K. Darrow
Bell Telephone Laboratories, Incorporated, New York

1. Introduction

THE major subjects of this article are two extensions of what formerly was called *atomic theory*—that is to say, the attempt to explain as many as possible of the properties of pieces of matter large enough to be visible and tangible and ponderable, by visualizing these as swarms of tiny particles each endowed with only a very few and simple qualities. Among the properties of gases, for example, are pressure and viscosity and entropy and temperature. Conceivably one might invest the ultimate atoms with all four. The atomic theory of gases as it stands today, however, is the outcome of a very different procedure. It is the achievement of an effort to interpret these four properties and several more as features of an imagined assemblage of enormously many corpuscles all alike, and not possessing them nor any others except position and velocity and mass (and moment of inertia, sometimes) and the liability to make elastic impacts with each other. On the whole the effort has been remarkably successful. Therefore viscosity and temperature and entropy are not attributed to single atoms, but pictures and expressions for them are derived as qualities of the assemblage. The theory which leads to these results is called *statistical;* it is based on certain assumptions which, in the form in which they were originally made, constitute the *classical statistics*. The successes of the classical statistics are a part of the evidence that matter is corpuscular. Once they were nearly the whole of the evidence, for they antidated the striking demonstrations of individual atoms which now spring to the mind whenever one is asked to state the reasons for accepting the atomic theory.

Radiation resembles a gas in some respects. Entropy and temperature and pressure, for example, are properties displayed by radiation when enclosed in a space surrounded by a wall of even temperature, just as they are by a gas in a like situation. It seems quite natural that one should try to interpret them in the same way as for a gas they are interpreted by the atomic theory: imagining the radiation as an assemblage of innumerable particles, a swarm of photons or corpuscles of light. Nowadays at least the idea seems quite natural; but of course, in the years when no one as yet had broken away from the tradition that light is altogether wavelike, it would doubtless have been thought a very wild one. Even after Einstein had ventured such a breach with the past, nearly a score of years elapsed before there was developed out of the theory of quanta an adequate conception of the "radiation-gas." The historical sequence in the growth of the atomic theory of matter was here

inverted: there was abundant evidence for the corpuscular theory of light, in phenomena such as the Compton effect and the photoelectric effect showing the work of individual photons, before the statistical theory of these corpuscles was perfected. We now see that the trouble was, that even when one accepts the notion of corpuscles of light without reserve, and even when one knows the proper values of energy and momentum to be assigned to these corpuscles, it still is not correct to apply to them the same statistics as gives such good results when applied to the atoms of matter. Bose discovered how to remodel the statistics, in order to construct a competent atomic theory of the radiation in thermal equilibrium in an enclosure.

The other of the new extensions of atomic theory is partly a revival — the resurrection of the theory, first proposed some thirty years ago, that part at least of the negative electricity within a metal acts like a swarm of freely-flying corpuscles which collide now and again not with each other but with the atoms. It was of course the classical statistics which was always used in developing this theory. Moribund because of several incurable discordances with fact, the theory was resuscitated by Pauli and by Sommerfeld with a revision of the statistics. It was not quite the same revision as enabled Bose to set up an atomic theory of radiation, but a very similar one, invented first by Fermi and later independently by Dirac. One cannot say that the so-renovated "electron-gas theory" is a perfect explanation of all the multifarious phenomena of the flow of electricity and heat inside of metals and outward through the boundaries of metals. Its initial successes, however, are so auspicious as to suggest that the hope of further progress lies not in renouncing it (as seemed to be almost inevitable before the alterations) but in amending it in its details.

Is the atomic theory of material gases to remain untouched by these novel ideas? Apparently all three forms of statistics, the classical and the two recent types, lead to very nearly the same conclusions when applied to material gases. Only at remarkably low temperatures and remarkably high densities do their predictions diverge; and under these conditions the experimental data are not easy to interpret for that purpose. Suppose, however, that eventually the data are proved to decide for one of the new forms of statistics against the old: what then? Probably we shall merely remove one of the theoretical foundation-stones of the kinetic theory of gases and insert another to take its place, meanwhile leaving practically intact the great superstructure of formulae and equations whereby the kinetic theory makes contact with experience. Happily this is an easier process in theoretical physics than in architecture.

Custom has lately changed the meaning of the term *atomic theory*, making it almost synonymous with *theory of the structure of the atom;* but the province which this latter has taken for its own is one to which its forerunner disclaimed all right of entry. It was never supposed that all of the properties of a gas can be interpreted as statistical features of a swarm of corpuscles. The earlier atomic theory conceded some of them to the individual atoms, thus in effect renouncing the ambition to explain them; and

among these were the spectra. Where the statistical theory left off, the builders of atom models took up the work. Bohr, for example, designed a model for the individual hydrogen atom, competent—at least to a great extent—to explain the Balmer series and the rest of the line-spectrum of "atomic hydrogen." This model he constructed, following Rutherford, out of a pair of corpuscles. What he and his successors thus developed was in a way an *atomic theory of the atom*—a degree deeper, or further, or higher perhaps, than the atomic theory of matter which had provided him with the notion and the scale of the atom to begin with.

What then distinguishes this new "atomic theory of the atom" fromt is ancestor? Well, the major differences in method and in aim are traceable to the fact, that in the later theory the number of elementary particles which constitute the system is quite manageably small, while in the earlier, it is inconceivably tremendous.

Bohr constructed his model for the hydrogen atom with only a pair of corpuscles, and those for all the other atoms out of not more than a few dozen each. Now with a model consisting only of two particles, one can specify positions and velocities for these with the utmost of precision, and go merrily ahead predicting and describing orbits with as much exactness as one cares to lavish. Even with dozens of electrons and a nucleus one can attain at least a specious accuracy of detail; remember the portraits of the electron-orbits of massive atoms which six or seven years ago were so profuse. Perhaps it is not wise to make such definite assertions; but it is feasible. Not so, however, with the subjects of the older theory.

The model proposed for a cubic centimetre of gas under the ordinary conditions of temperature and pressure consists of something like 10^{20} particles. Merely the mention of so extravagant a figure is sufficient to persuade that it is vain to dream of making any progress by postulating a definite position and a definite velocity for each of these. The life of the human race would not be long enough to write down even the postulates, to say nothing of the inferences.

This seems a fearful handicap; but it is not so at all. Adopting the statistical method, one does not even begin upon the hopeless task of fixing place and motion for every particle. We content ourselves with writing down a function, which states how many among the multitude of particles we assume to be situated in each small (but not too small) element of volume; and how many we assume to have momenta which lie in each small (but not too small) range of momentum. These are specifications much more modest and vague; but they are ample. For the things which we wish to interpret—entropy and temperature, viscosity and conduction and diffusion—the atomic picture need not be made one whit more definite.

In saying this I am understating the case. If the atomic picture could be made more definite, say by stating the locations and the velocities of all the atoms with absolute precision, the meanings which we shall presently attach to entropy and temperature would be dissolved. Our theory of these entities depends upon the vagueness of the picture. Seemingly they appeal

to us as physical realities because our senses and our instruments are too obtuse to perceive the atoms. Our minds must feign a somewhat similar obtuseness, pretending not to fix the particles of the imagined swarm too sharply; therefore it does not matter that they are so numerous that the pretence becomes sincere. Exact knowledge of the individual atoms is unattainable; but it is useless, is not desirable even. One remembers Aesop's tale of the fox and the inaccessible grapes; in this case it is probable that the grapes really *are* sour.

For that matter, perhaps they do not even exist. One of the most striking of the very recent ideas in theoretical physics is the thought, that even for atom-models with but a few particles, even in thinking of an isolated particle, it may be altogether pointless to assign exact positions and velocities. In dealing with a swarm of particles by the statistical way, we do in effect fix the position of each corpuscle, but with a certain latitude; we fix the velocity of each, but again with a certain latitude. Perhaps this latitude, this indefiniteness, is something inherent in nature. Insisting as I am upon the contrast between the theory of the structure of the atom and the statistical theory of matter and radiation, I may in effect be insisting on the contrast between a faulty way of visualizing some phenomena, and a correct way of visualizing all.

A function of the sort which I just mentioned, a so-called *distribution-unction*, is the goal of every statistical theory. I have said that it states how many among the multitude of particles we *assume* to be located in each small element of space and to have momentum comprised in each small range of values of momentum. So it does; but the purpose of a statistical theory is, to derive it from assumptions still more fundamental, in preference to assuming it outright. Of course one might say instead, that the reason for deriving a distribution-function is to put the fundamental assumptions to their test. Whichever viewpoint one prefers, it is the distribution-function which is tested by experiment: indirectly, in that it supplies numerical values for such things as conductivity, viscosity, specific heat; and directly, for there are now immediate ways of observing it in certain cases.

A distribution-function commonly appears in an equation of this form:

$$dN = f(x, y, z, p_x, p_y, p_z) \cdot dx dy dz dp_x dp_y dp_z. \qquad (1)$$

Such an equation will as a rule refer to some particular assemblage of particles, say N altogether, occupying some definite region of space: a gas in a tube, radiation in a cavity, electrons in a wire. It is to be read as follows: "dN, equal to $f.\ dxdydzdp_xdp_ydp_z$, stands for the number of particles having coordinates in dx at x, in dy at y, in dz at z, and components of momentum in dp_x at p_x, in dp_y at p_y, in dp_z at p_z." The phrasing "in dx at x" is a succinct alternative for "between x and $x+dx$."

The function f is the distribution-function in the variables in question—here the *coordinates* of the particles referred to some Cartesian frame in the ordinary or "coordinate" space, and the components of momentum resolved along the axes of that frame, the *momenta*. Heretofore it has been

customary to use the components of velocity rather than the momenta, but these are much to be preferred: partly because it is they which figure in the canonical equations, but chiefly because we shall find when we pass over to the study of assemblages of photons that the momenta play the same role in these as they do in assemblages of atoms, while the speeds of all photons are the same. There is a well-known formula for translating a distribution-function from one set of variables to another set dependent on the first, which we shall use in special cases.[1] It is also well known that to obtain the distribution-function in *some* of the independent variables from the distribution-function in *all* of them, it is necessary to integrate the latter over the entire range of all the *other* variables: in such a case as is symbolized by equation (1), the distribution in p_z would be obtained by integrating f with respect to the first five variables over the entire range of each.

The product $dxdydz$ is an element of volume in ordinary or *coordinate-space*; the product $dp_x dp_y dp_z$ is an element of volume in *momentum-space*, in which each particle is represented by a point having for its coordinates in a Cartesian frame the values of its momenta; the product $dxdydzdp_x dp_y dp_z$ is an element of volume in *phase-space*. The function f describes the distribution of the assemblage in this phase-space of six dimensions. In some cases—for instance, that of electrons in a metal not at an even temperature, and that of oscillators—we shall have to think continually of this six-dimensional space. In others—whenever we deal with photons, and whenever we consider atoms or electrons in a region where neither temperature nor potential varies from place to place—we shall be able to assume that the distribution in the coordinate-space is uniform (that f is independent of x, y, z) and to dismiss it from mind, and to derive the distribution in the three-dimensional momentum-space quite separately as if there were no other. Even in these simplest cases it would no doubt be more consistent to operate always in the phase-space. Unhappily the human mind is so constructed, that no matter how much it may ratiocinate about space of six dimensions or six trillion, it always visualizes in space of three.

In an equation such as (1), the differential element or the product of such elements which terminates the right-hand member must be neither too large nor too small. If it is so large that f varies considerably from one point in it to another, then its multiplier, which is by definition the *mean value* of f in the said element, must be computed by the methods of integral calculus. If on the other hand it is so small that it contains only a few of the corpuscles, then the product of f into its size may be many times as great or many times as small as the number which it does contain. This is easily

[1] Let $u_1, u_2, \ldots$ represent the variables of the first set, $v_1, v_2, \ldots$ those of the second; let $f(u_1, u_2, \ldots)$ and $F(v_1, v_2 \ldots)$ stand for the distribution-functions in the two sets; then

$$F(v_1, v_2, \cdots) = f(v_1, v_2, \cdots \frac{\partial(u_1, u_2, \cdots)}{\partial(v_1, v_2, \cdots)}$$

perceived by proceeding to the absurd limit of dividing the space into say ten times as many elements as there are corpuscles, so that in at least nine-tenths of the elements the number of participles is zero while f is greater than zero, and in the others the number is generally much greater than f times the size of the element. To subdivide the space so finely would be to make the atomic picture too definite, and ruin it for the purposes for which we now require it.

It is not too early in this paper for me to say emphatically that the differential elements which figure in equations such as (1) *must not be identified* with the elementary compartments of the phase-space, which we shall presently encounter, and which are so important in the new statistics and in the old alike. It takes a great many of these latter to make up an element large enough to be employed in an equation like (1). Otherwise expressed: the subdivision of the phase-space into the elementary cells or compartments of the forthcoming theory is much too fine to be used in connection with the distribution-function. Much confusion may arise from failing to realize this.[2]

In speaking of the distribution-function, I have been tacitly assuming that there is such a thing as a stable, self-sustaining, changeless distribution of the atoms of a gas, the photons in a cavity, the electrons in a wire. This assumption must now be examined. It is scarcely self-evident; one might guess at first that the more numerous the particles, the more abruptly would the distribution vary from one moment to the next, and that an assemblage of 10^{20} particles would be in such unceasing turmoil that it would be senseless to imagine one single distribution for it.

Experience however shows the reverse. The gas in a tube remains uniformly dense and stationary, it does not surge forever to and fro nor huddle in a corner nor become spontaneously hot at one end and cold at the other. In the radiation in a cavity with heated walls the intensity comprised within any portion of the spectral range remains unchanged so long as the temperature of the walls is constant. The distribution-in-velocity of the electrons streaming from a heated filament does not appear to change. Moreover, when by artifice the gas in a tube is forced to assume uneven density, non-uniform temperature, or any sort of flow or turbulence, it settles down very quickly into a stagnant uniformity as soon as it is left to itself.

Now we know that while a gas is passing from an unstable state—a state of non-uniform temperature, for instance—to its stable and permanent condition, a property which we call its *entropy* and denote by S is increasing; in certain simple cases we can evaluate this rate of change of entropy. We know that when a gas is in its stable condition, its entropy is at a maximum;

[2] I am thinking particularly of the fact that in most expositions of the classical statistics one is adjured that there must be many particles in each compartment, and then in taking up the Fermi statistics one is told that there must be not more than one in each compartment; yet the two lead to formulae which in the limiting case are the same.

we know how to compute the entropy (except perhaps for an additive constant) of a given quantity of a gas in this condition, as a function of its temperature and others of its measurable properties. And when we have evaluated both the entropy S and the energy E of a gas under any specific conditions, we know that its absolute temperature is determined by the following equation,

$$dS/dE = 1/T \tag{2}$$

which is the definition of absolute temperature.

If we had obtained by some independent way an adequate atomic picture of entropy, so that whenever a distribution-function was suggested we could compute the value of S: then necessarily the stable distribution would be the one for which S has the greatest value compatible with the given number of particles and the given amount of energy. We do not have an independent way. But if instead we adopt some *tentative* atomic picture of entropy, some function S of which we can compute the value for any given distribution: then the test of our picture will be, whether the distribution for which this tentative S has its greatest value is verified by experiment to be the stable one. It will be found that this distribution "of maximum S" involves the derivative dS/dE, and therefore the absolute temperature; so the temperature enters into the postulated distribution-function in the course of its derivation, not by separate assumption or by an afterthought.

This method is the very notable one invented by Boltzmann, and continued by Planck. One choice of the function S which is to be identified with entropy leads to the classical or Maxwell-Boltzmann distribution-law; another leads either to the Bose or to the Fermi distribution, the difference between these two entering in at another point.

Each of these suggested functions is logarithmic; it is proportional to the logarithm of a function which is called *probability*. In theoretical physics it is a fairly general rule, that when a theorist introduces the word *probability* he is abandoning all hope of explaining by cause-and-effect the phenomena of which he is discoursing. This is the disadvantage of Boltzmann's method. The "distribution of maximum S" is baptized "the most probable distribution;" there is even a numerical estimate of its "probability," and in general it turns out to have so much greater a probability than all the others put together that one accepts without demur the conclusions that in practice it will be stable. But there is no proof that the "most probable" distribution is always or even usually followed by another exactly like it, nor that an "improbable" distribution is always or even usually followed by another of greater probability; there is no study of the way in which one distribution is transformed into another, there are no assumptions about the collisions or encounters which presumably offer to the particles their means of interchanging speed and energy, and to the assemblage its means of approaching the stable distribution. There are other statistical methods in which account is taken of these things, and we shall have a glimpse of one of them in the last section of this paper; but the notion of

causality is absent from the method which will be followed in deriving the distribution-laws of Maxwell and Boltzmann, of Bose, and of Fermi and Dirac.

These three distribution-laws will be applied to freely-flying particles in regions which are either field-free, or else pervaded by a field (electrostatic or gravitational) derivable from a potential. It may surprise the reader to hear so little about oscillators, considering that the statistics which Planck applied to these objects was the first of all the modifications of the classical statistics, was the source of the entire quantum-theory, and therefore the most important advance of the physics of the last quarter-century. The history of this period is very curious; but I cannot mention more than a couple of the salient points.

The Planckian oscillators served two purposes: they enabled Planck to derive the law of distribution of radiant energy at uniform temperature in a cavity, by supposing the radiation to be entirely wavelike and to be in equilibrium with myriads of oscillators in the walls of the cavity; and they enabled various savants to develop, step by step, a progressively improving theory of the specific heat of solids. The Bose statistics made them quite superfluous for the first purpose: by applying this statistics to the radiation supposed to consist of corpuscles, we can derive the same law of distribution without invoking the oscillators at all. As for the second: as early as 1912 (which seems remarkable, now) Debye had replaced the concept of a solid as a latticework of vibrating atoms by the concept of a solid as a system of stationary waves agitating a continuum. I do not mean to imply, of course, that the existence of the atoms was denied; I mean no more than to say, that in these statistical reasonings the individual vibrating atom was replaced by an individual pattern of stationary waves. Today we are becoming familiar with the idea that in certain reasonings, a freely-flying electron or quantum or even an atom in a region bounded by walls may be replaced by a pattern of stationary waves filling the whole of this region. Thus it seems that the free particle, the oscillator, the stationary wave-pattern, are in close affinity with one another; they may simply represent different ways of looking at the same thing. Though on almost every page of this article I shall write in the language of the strictest corpuscular theory, it is probable that every one of the results could be translated into the language of oscillators or the language of waves.

There are still assumptions to be made about the individual particles. They are to have position, momentum and energy. Momentum may be separated into mass and velocity; in many ways it is better left as an elementary concept. It will turn out that the essential difference between photons on the one hand, electrons and atoms on the other—that is to say, the essential difference between the particles out of which we shall try to build a picture of radiation, and those of which we shall build models of gases and of electricity in metals—lies in the relation between momentum and energy.

Experience with matter in bulk leads to the well-known equations connecting kinetic energy K and momentum p with mass m and speed v:

$$K = \tfrac{1}{2}mv^2, \qquad p = mv \tag{3}$$

and these are supposed to hold for the ultimate particles of matter and of electricity. During the years in which the corpuscular theory of light was struggling into existence—for, it will be remembered, light was still considered to be entirely wavelike even after Planck had founded the quantum-theory by his statistics of oscillators—Einstein proposed at two different times (1905 and 1917) the following formulae for the energy and the momentum of photons in terms of their wavelength:

$$E = hc/\lambda, \qquad p = h/\lambda. \tag{4}$$

Historically it is interesting that he proposed the latter formula because of certain statistical studies which he had made of the equilibrium between photons and atoms. The verification of the latter by the Compton effect, of the former by the photoelectric effect and many other phenomena, is too familiar to require comment.

Now from equations (3) we deduce, for particles of matter and of electricity:

$$E = \frac{1}{2m}p^2 = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2) \tag{5}$$

and from equations (4) we deduce, for particles of light:

$$E = pc. \tag{6}$$

The difference between these two relations is responsible for some of the differences between radiation-gas on the one-hand, electron-gas and material gases on the other; but by no means for the major part. The major difference lies in the statistical theory as we shall now find out.

The Classical Statistics

We are going to represent three kinds of objects—ordinary or material gases, radiation in enclosures, negative electricity in metals—as assemblages of particles possessing location and momentum. We may visualize such an assemblage first as a swarm of points in ordinary space, with a coordinate-frame along the axes of which the coordinates x, y, z of the particles are measured; then as a swarm of points in momentum-space with a frame along the axes of which the momenta p_x, p_y, p_z are measured.

I will first illustrate the method of classical statistics by using it to ascertain the most likely distribution of particles in ordinary space, a case where seemingly the result may be foreseen. For it seems a truth of intuition that inside a box of ordinary space, with nothing (*e.g.* no variations of potential) to distinguish one region from another, the particles must tend to distribute themselves uniformly. This is a conclusion to which the statistical method *must* lead. The uniform distribution *must* be the most probable.

How then should we define the "probability" of a distribution so that it shall be greatest for the uniform one?

But in the first place, what *is* a uniform distribution? We must divide the space—mentally, of course—into compartments of equal volume. The distribution will then be called uniform, if the numbers of particles in the various compartments are about the same. But this clearly requires that these subdivisions be of a certain size. Their linear dimensions cannot for example be smaller than the average distance between particles, as then a "uniform distribution" would be impossible. To partition the space too finely would be like studying a painting with a microscope. The quality which we wish to define evades too sharp a scrutiny. The compartments should contain large numbers of particles, both for the stated reason and for the convenience of a certain mathematical approximation which is made.

Denote then by N the total number of particles, by m the number of compartments into which the volume V is divided, by N_i the number of particles in the ith compartment. A distribution is described by stating all the numbers $N_1, N_2 \ldots, N_i, \ldots N_m$.

The basis of the classical statistics is the fact that if the particles have identities—if each of them is labelled by a distinctive letter, for instance—there are different ways of arranging them in the same distribution. One starts with any arrangement compatible with the prescribed "populations" $N_1, N_2, \ldots N_m$, and obtains all the other arrangements by interchanging particles *ad libitum* among the compartments, respecting only the condition that each of these shall always have as many as it had at first. The total number of distinct arrangements, the number of *permutations of the combination* $N_1, N_2, \ldots N_m$, is by a well-known theorem:[3]

$$W = \frac{N!}{N_1! N_2! \cdots N_m!} \tag{7}$$

This number has its minimum value of unity for a distribution in which all the particles are crowded into one compartment, which would be the most non-uniform conceivable; and its maximum value for the uniform distribution, as I now proceed to show.[4]

[3] Imagine yourself stationed beside a set of m baskets and an urn filled with N lettered but otherwise indistinguishable balls, which are to be lifted out at random and dropped into the baskets under the following rules of the game: the first N_1 which come to your hand are to be dropped into basket 1, the next N_2 to come to your hand are to go into basket 2, and so on to the end. Having acted accordingly, you note down the assortments of balls in the various baskets, and repeat the process *ad infinitum*. Now there are $N!$ different orders in which the balls may come out of the urn. When the inspection of the baskets after two drawings reveals different results, the orders must certainly have been different. But two different orders need not reveal two different results to the inspection. Take any order, to start with; then there are $(Q-1) = (N_1! N_2! \cdots N_m! - 1)$ others which yield the same result. For there are $N_1!$ orders in which the earliest N_1 balls to emerge might come out, without any of them losing its place among the first N_1; there are $N_2!$ orders in which the next N_2 might come, without any losing its place in the second basket; and so forth. Each of the $N!$ orders then is but one among Q altogether which lead to the same result; so that there are only $N!/Q$ different results.

[4] What will actually be shown is that for the uniform distribution the function W is stationary; that it is maximum (not minimum) seems fairly obvious from the physics of the case, but can be proved.

Let us use the logarithm of W instead of W itself. If W has a maximum for any distribution so also will its logarithm, which is easier to handle, and will presently be chosen as the representation of entropy. We have:

$$\log W = \log N! - \sum_i \log N_i! \tag{8}$$

Now we introduce Stirling's approximation for the factorial of a large number—by far the greatest and the most frequently invoked of the mathematical aids in statistical theory. It is:

$$\begin{gathered} x! = (2\pi x)^{1/2}(x/e)^x \\ \log x! = x \log x - x + \tfrac{1}{2} \log (2\pi x) \end{gathered} \tag{9}$$

The first two terms of this latter expression form an approximation singularly good even when x is no greater than ten or thereabouts. Using it we have:

$$\log W = \text{const.} - \sum N_i \log N_i. \tag{10}$$

Denote by W^0 the value of W for some particular distribution $N_1{}^0$, $N_2{}^0, \ldots N_m{}^0$ and by $W = W^0 + \delta W$ its value for some other only slightly different distribution $N_1{}^0 + \delta N_1$, $N_2{}^0 + \delta N_2, \ldots N_m{}^0 + \delta N_m$. The difference between the values of log W for these two distributions is to first order of approximation:

$$\delta \log W = \delta W / W = - \sum_i (1 + \log N_i{}^0) \delta N_i. \tag{11}$$

If W^0 is a maximum for the distribution $N_1{}^0 \ldots N_m{}^0$, then the difference between log W^0 and the value of log W for any other slightly different or "slightly varied" distribution must vanish to first approximation. The quantity on the right of (11), the "first variation" of log W, must be zero for any permitted set of values of $\delta N_1, \ldots \delta N_m$; meaning by "permitted" any set of integer values adding up to zero, for we consider an assemblage of an invariable number of particles.

Now one sees immediately that the right-hand side of (11) does vanish, if all the populations $N_i{}^0$ have the same value, say α; for then

$$\log W = - \sum_i (1 + \log \alpha) \delta N_i = \text{const.} \sum_i \delta N_i \tag{12}$$

and the permitted variations are precisely those, for which the summation $\sum N_i$ is zero.

We do therefore reach the result which was desired. Failing it, this mode of "counting the ways in which a distribution may be realized" would have been unprofitable. As it is, the quantities W and log W are greatest for the uniform distribution which seems intuitively the most probable and is the rule for gases, and least for the utterly non-uniform one which seems the least probable. Tentatively the former is adopted as measure of the "probability" of a distribution.

I point out in passing that while the foregoing result is mathematically valid for any value of the constant α, the total number of particles prescribed for the assemblage determines the value of α which is physically permissible: viz. N/m.

We proceed to apply this method to the swarm of points in momentum-space representing the assemblage.

Like the coordinate-space, the momentum-space is to be divided into equal compartments large enough to contain each a multitude of particles. We are to define a distribution by specifying how many particles are in each compartment, and calculate as before the number W which is to measure the "probability" of the distribution. The values for W, for log W and for the variation of log W are obtained just as before. There is however an important novelty. Since the energy of a particle depends on its position in momentum-space, different distributions usually entail different values for the total energy of the assemblage. If we compute the variation of log W due to a slight change in distribution, we shall usually be computing a variation in log W correlated with a certain variation of the total energy U of the assemblage.

We now take the very great step of *identifying the quantity log W with entropy.*

More precisely, we assume that the entropy S is proportional to the logarithm of W:

$$S = k \log W \tag{13}$$

introducing a constant factor k, and relying on subsequent experiments to teach us its numerical value.

Now when a gas being initially in thermal equilibrium at temperature T receives an infinitesimal amount of energy dE, and regains thermal equilibrium with its augmented energy, its entropy ascends by the amount of dS given by the equation:

$$dS/dE = 1/T. \tag{14}$$

If then the foregoing model of the gas and the foregoing picture of entropy are justified, the variation of log W in passing from the most probable distribution consonant with a total energy E to the most probable distribution consonant with a total energy $E+dE$, (the total number of particles remaining the same) must be equal to $(1/kT)dE$.

If we start from the most probable distribution for energy E and make *any* slight change in it involving an energy-change dE, the new distribution will presumably differ but little from the most probable distribution for $E+dE$. We therefore say: the most probable distribution for energy E is the one of which the first variation is dE/kT. This expression vanishes, if we are comparing distributions for which E is the same; which is as it should be.

It is now easily shown that such a distribution is the following

$$N_i = \alpha \exp(-\epsilon_i/kT) \tag{15}$$

in which α stands for any constant and ϵ_i for the average energy of particles in the ith compartment, which is related to the average momenta of these particles by the equation

$$\epsilon = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2) \tag{16}$$

for we have only to write down the expression for δS as furnished by equation (11), and introduce into it the value of $\log N_i$ as supplied by equation (15):

$$\delta S = -k \sum(1+\log N_i)\delta N_i = -k(1+\log \alpha) \sum \delta N_i + \sum \epsilon_i \delta N_i / T = \delta E / T \tag{17}$$

the result which was desired.

The value to be chosen for the constant α will be determined as before by the total number of particles. Denote this number by N, and conceive the compartments as tiny cubes of volume H, so that there are $1/H$ of them per unit volume of the momentum-space. The density ρ of the particles in momentum-space, which is no other than the *distribution-function in the momenta*, is given anywhere by the value of N_i/H computed for the value of energy there prevailing:

$$\rho = N_i/H = \frac{\alpha}{H} \exp(-\epsilon/kT) \tag{18}$$

$$= \frac{\alpha}{H} e^{-p_x^2/2mkT} e^{-p_y^2/2mkT} e^{-p_z^2/2mkT}$$

and it is the integral of this expression over the whole of momentum-space which is equal to N:

$$N = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \rho \, dp_x dp_y dp_z. \tag{19}$$

The integration is easily affected; the triple integral is the product of three identical single integrals, and we have:

$$N = \frac{\alpha}{H}\left[(2mkT)^{1/2} \int_0^{\infty} e^{-w^2} dw\right]^3 \tag{20}$$

w being a symbol for each of the three momenta in turn; so that

$$\alpha = \frac{NH}{(2\pi mkT)^{3/2}}. \tag{21}$$

The expression for the number of particles in any compartment thus becomes:

$$N_i = \frac{NH}{(2\pi mkT)^{3/2}} \exp(-\epsilon_i/kT) \tag{22}$$

involving the four constants m, N, k and H. The first three are determinable by experiment, in ways which I will later mention; the third is the universal constant known in Boltzmann's honor by his name, though he himself never evaluated it. The fourth, the volume II assigned to the compartments, drops out of the distribution-functions—out of the function ρ, out of the distribution-in-energy soon to be deduced, out of the fundamental distribution-function f in the coordinates and momenta defined by equation (1), and which I now set down in place of ρ:

$$f=\frac{N}{V(2\pi mkT)^{3/2}}\exp\left(-\frac{p_x^2+p_y^2+p_z^2}{2mkT}\right) \tag{23}$$

V standing for the volume in ordinary space of the enclosure which contains the assemblage. This evasion of II is very deceptive; for its suggests not merely that the exact volume of the compartments is of no importance, but that the compartments themselves were invented only as a momentary stepping-stone to the distribution-functions, and should be allowed to shrink to zero like the infinitesimals of the calculus. This however is precisely what is not allowed. It is of the essence of the argument that there are compartments of finite size. As will presently transpire, I suspect that the division of momentum-space into compartments should be regarded as a quantum postulate, even in this case of the derivation of the Maxwell-Boltzmann law which seems to be at the opposite extreme from all the notions of quantum-theory.

The next step is the derivation of the *distribution-in-energy*. In preface I point out that the distribution which we are considering is, in respect to the directions of motion of the particles in ordinary space, *isotropic*. Mathematically, this occurs because p_x, p_y and p_z enter symmetrically into all the distribution functions; physically it occurs because we have made no assumption leading to a preference of any direction over any other. Later on we may establish a preferred direction by introducing a field of force, and then the impending steps may have to be reconsidered. Until then the distribution which we shall study will be described completely by saying that they are isotropic and giving the distribution-function-in-energy. This may be obtained from the distribution-function-in-the-momenta by transforming to a polar coordinate frame in the momentum-space.[5]

I follow practically the same route.

Divide up the momentum-space into spherical "shells" by means of a sequence of spheres all centered at the origin. Each sphere corresponds to a

[5] Denote by p the quantity $(p_x^2+p_y^2+p_z^2)^{1/2}$ which is the magnitude of the momentum; and by θ and ϕ the angles which with p constitute a spherical coordinate system. We have

$$\rho dp_x dp_y dp_z = \rho p^2 \sin\theta d\theta d\phi dp = \frac{\alpha}{H} e^{-p^2/2mkT} p^2 \sin\theta\, d\theta\, d\phi\, dp$$

and the distribution-in-momentum is obtained by integrating over all values of θ and ϕ, the distribution-in-energy from it by means of the relation (5).

value of ϵ, each shell to a range $d\epsilon$ of values of ϵ. Take one of the latter at random; call it shell s, denote by ϵ_s and by ϵ_{s+1} or $\epsilon_s+d\epsilon$ the energy-values at its boundary spheres, by r_s and r_s+dr the radii of these, by dV the volume of the shell. Then:

$$r_s=(2m\epsilon_s)^{1/2}, \qquad dr=\left(\frac{m}{2\epsilon_s}\right)^{1/2}d\epsilon$$

$$dV=4\pi r_s{}^2dr \qquad dr=\frac{2}{(\pi)^{1/2}}(2\pi m)^{3/2}(\epsilon_s)^{1/2}d\epsilon. \tag{24}$$

Suppose to begin with that each shell is large enough to contain very many compartments. The number Q_s of compartments in shell s will then be:

$$Q_s=dV/H=\frac{2}{H(\pi)^{1/2}}(2\pi m)^{3/2}\epsilon_s{}^{1/2}d\epsilon \tag{25}$$

and the average number of particles per compartment in shell s, call it N_s, will be:

$$N_s=\alpha\exp(-\epsilon_s/kT) \tag{26}$$

and the total number M_s of particles in the shell will be:

$$M_s=Q_sN_s=\frac{2N\pi}{(\pi kT)^{3/2}}\epsilon_s{}^{1/2}e^{-\epsilon_s/kT}d\epsilon=F(\epsilon_s)d\epsilon. \tag{27}$$

This is the number of particles having energy-values in $d\epsilon$ at ϵ_s. Hence the distribution-function-in-energy F is the factor multiplying $d\epsilon$ (it would be well to discard the subscript s in writing it). I have copied the value of α from (21), but it could have been derived by integrating F from $\epsilon=0$ to $\epsilon=\infty$ and equating the integral to N.

The separation of M_s or $F(\epsilon)d\epsilon$ into two factors—Q_s the number of compartments in the shell s, N_s the average number of particles per compartment—is highly advantageous in searching for the distinctions between the various proposed statistical laws. We shall see that in passing from one to another sometimes one of the factors is changed, sometimes the other, sometimes both.

In particular, we may pass from the Maxwell-Boltzmann law to a distribution like that which Planck derived for oscillators, simply by changing the factor Q_s. We have been dividing the momentum-space into compartments of equal volume, so that the number comprised in a shell s between spheres ϵ_s and $\epsilon_s+d\epsilon_s$ is proportional to $\epsilon_s{}^{1/2}d\epsilon_s$. Let us instead divide it into compartments of which the volumes increase steadily from the origin outward, at such a rate that the number in a shell s is proportional to $d\epsilon_s$ without the factor $\epsilon_s{}^{1/2}$.

This is, of course, not the way in which Planck's postulate is habitually stated, though it is substantially the way in which Planck stated it himself. Usually it is said, that Planck restricted the energy of the particles of the

assemblage to a set of "permitted values" spaced at equal intervals: say the values a, $a+b$, $a+2b$, $a+3b$,.... where a and b stand for constants. Each of these permitted values corresponds to a sphere in the momentum-space. In the shell s there are approximately $d\epsilon_s/b$ of these "permitted spheres"; the approximation being closer, the larger ϵ_s and $d\epsilon_s$ are in comparison to b. Now whether we conceive that the $d\epsilon_s/b$ sets of particles in the shell s are located on the surfaces of as many permitted spheres, or alternatively that they are scattered through as many compartments, is for the statistical results of no importance. There may be other reasons for preferring one picture to the other; but the predictions of the statistical theory are the same, whichever is adopted. I will therefore alternate between the two pictures, retaining for the moment that of a subdivision of the momentum-space into compartments; but now it will be expedient to think of these as thin spherical films, centered at the origin and increasing in volume from the innermost outward at the specified rate.

If the shell s is large enough to contain many of these compartments or permitted spheres, we may use the first approximation for the number which it contains:

$$Q_s = d\epsilon/b \tag{28}$$

and putting the expression (26) for the number of particles per compartment, we get:

$$M_s = Q_s N_s = \frac{\alpha}{b} \exp(-\epsilon_s/kT) d\epsilon_s = F(\epsilon_s) d\epsilon_s \tag{29}$$

for the number of particles having energy-values between ϵ_s and $\epsilon_s + d\epsilon_s$. The value of the constant is fixed as heretofore by the condition that the integral of F over the entire range of energy from 0 to ∞ shall be equal to N:

$$\int_0^\infty F(\epsilon) d\epsilon = \frac{\alpha kT}{b} \int_0^\infty e^{-w} dw = N \tag{30}$$

so that we arrive at the following distribution-in-energy function:

$$F(\epsilon) = \frac{N}{kT} \exp(-\epsilon/kT). \tag{31}$$

This function certainly does not display any feature which suggests the achievements of Planck! It looks as smooth and continuous as the Maxwell-Boltzmann function itself, and the constant b, the step or interval between the successive permitted energy-values or the boundaries of successive compartments, is nowhere to be seen. The constant b however has slipped out for the same reason as the constant H from the function (23), and the apparent continuity is due in both cases to the same cause. In preparing and effecting the integration (30) in order to obtain a value for the constant α, we assumed that the various permitted energy-values within

the range $d\epsilon_s$ are all sufficiently nearly equal to be identified with the single value ϵ_s. That is to say, we smoothed over the discontinuities which had previously been brought in by the assumption of separate compartments. No wonder that there is not a sign of them in the function (31), even as there is not a sign of them in the Maxwell-Boltzmann law!

We might however avoid this smoothing-over, if we could attain the value of α by an actual summation over the various compartments instead of by integration. Now with Planck's postulate this is mathematically feasible and indeed easy. For the number of particles in the ith compartment being

$$N_i = \alpha \exp(-\epsilon_i/kT) \tag{32}$$

the total number of particles is computed thus:

$$N = \sum N_i = \alpha e^{-a/kT} \sum_{i=0}^{\infty} e^{ib/kT}$$

$$= \frac{\alpha e^{-a/kT}}{1 - e^{-b/kT}} \tag{33}$$

by virtue of the very convenient consequence of the binomial theorem that $(1+x+x^2+\ldots.) = (1-x)^{-1}$; so that for α we obtain the *exact* value:

$$\alpha = Ne^{(a-b)/kT}(e^{b/kT} - 1) \tag{34}$$

and for the populations of the various compartments, the formula:

$$N_i = Ne^{-b/kT}e^{-ib/kT}. \tag{35}$$

Here the discontinuity implied in the classical picture of a momentum-space divided into compartments is admitted and accepted, as it never was in the process of deriving the Maxwell-Boltzmann law. Planck did not put discontinuity into the classical statistics; it was there already; he refrained from disregarding it. Instead of confining his studies to the circumstances in which it can safely be ignored, he extended them to ranges where it had to be taken account of, and he took account of it.

As I intimated, the distribution (35) was proposed by Planck not for freely-moving particles, but for oscillators. The "Planckian oscillator" may be visualized as a particle which executes simple-harmonic vibrations back and forth in a straight line across a position of equilibrium, to which it is attracted by a force proportional to its displacement. It is like a free particle, in that its state at any moment is described by giving the values of its position q and momentum p, q being measured from its point of equilibrium; but it is unlike a free particle in that its energy depends not on p alone but on both p and q, being a function of the form (Ap^2+Bq^2). Therefore we must envisage not the momentum-space alone but the phase-space of the variables p and q. In principle it would have been better, had we envisaged the phase-space all along; but since for an assemblage of free particles that space has six dimensions, it was impractical to visualize more than

the momentum-space, and since the energy depended only on the momenta that compromise was not detrimental except for one feature which I can later introduce. Here the compromise would be ruinous, but it is unnecessary since the phase-space has only two dimensions.

Visualize then this two-dimensional phase-space as a plane with p and q axes at right angles to each other. Suppose all the oscillators to have the same mass and the same natural frequency, which is to say, the same values of the constants A and B in the above-mentioned formula for their energy; but let them differ in amplitude. The point representing any oscillator in the phase-space runs round and round in an elliptical orbit centered at the origin. Different amplitudes correspond to different ellipses. The energy of an oscillator depends on its amplitude; therefore different energy-values correspond to different ellipses, and reversely. If we divide the phase-space into compartments by a succession of ellipses centered at the origin, each of these compartments corresponds to a specific range of energy-values. If the dividing ellipses are so spaced that these compartments are of *equal* area (equal volume of the phase-space), they correspond to *equal* ranges of energy-values—an important difference between this case and the one which was previously treated.

If the dividing ellipses are spaced to form equal compartments, they themselves correspond to energy-values forming a linear sequence: call these a, $a+b$, $a+2b$,....$a+ib$....as before. Whether we call these the "permitted" energy-values and allow the oscillators only the choice among them, or whether we sprinkle the oscillators uniformly through the compartments, makes only a secondary difference. In this case, in fact, we can easily see exactly what difference it makes. If the oscillators are sprinkled uniformly in each compartment, then by applying the classical statistics we get just the same distribution (35) as when we assume them restricted to the energy-values $(a+ib)$. But when we undertake to evaluate the average energy of all the oscillators, then in the one case we must put down the mean energy of those in the ith compartment as the arithmetic mean of the values $a+ib$ and $a+(i+1)b$, while in the other case we must put down the energy of those at the ith permitted ellipse as $a+ib$. Hence to change over from the picture of permitted energy-values to the picture of compartments is the same thing as to replace the original sequence of permitted energy-values by another sequence of values located midway between them. I mention this chiefly in order to emphasize that the subdivision of phase-space into compartments is *ipso facto* quantum-theory.

As every reader knows, Planck postulated that the quantity b—the interval between the permitted energy-values, or the energy-range within a compartment, whichever picture is chosen—is the product of a universal constant (h) and the frequency of the oscillators (ν). The area of the equal compartments is then equal to the universal constant[6] whatever the fre-

[6] The point in the phase-space representing an oscillator of mass m, frequency ν, and amplitude C describes an ellipse having semi-axes C and $2\pi m\nu C$ and area $2\pi^2 m\nu C^2$; its energy is $U = 2\pi^2 m\nu^2 C^2$; hence the relation between energy U and area F is

quency of the oscillators. From this latter statement the general principle is derived: To state it one must first adopt a symbol (say n) and a name (say *number of degrees of freedom*) for the number of distinct coordinates q required to describe the individual member of whatever assemblage one may be considering; this is also the number of distinct *momenta* p, there being one p for each q. Then the principle generalized out of Planck's postulate for oscillators is this: *For an assemblage of individuals with n degrees of freedom the phase-space is to be divided into compartments of volume* h^n.

We will now see what the classical statistics, supplemented by this principle, proposes for an assemblage of particles for which the relation between energy and momentum is $\epsilon = cp$ as it is for corpuscles of light, instead of $\epsilon = p^2/2m$ as it is for corpuscles of matter.

Different energy-values correspond as before to different spheres all centred at the origin of the momentum-space, but the numerical relations are changed. Instead of equations (24), we have:

$$\begin{aligned} \epsilon_s &= cr_s, \qquad d\epsilon_s = c\,dr_s \\ dV &= 4\pi r_s^2 dr_s = (4\pi/c^3)\epsilon_s^2 d\epsilon_s \end{aligned} \tag{36}$$

dV standing for the volume of the shell s covering the energy-range between ϵ_s and $\epsilon_s + d\epsilon_s$. Divide the momentum-space into compartments of equal volume H. We derive the "smoothed-over" distribution-function for the case in which N_i varies so little from one compartment to the next that even when the shell s is thick enough to comprise very many compartments the values of N_i for all of them may be equated to a mean value N_s. Under these conditions we may write for the number of compartments in the shell s,

$$Q_s = dV/H = (4\pi/c^3 H)\epsilon_s^2 d\epsilon_s \tag{37}$$

Putting down the classical value (15) for the number of particles in any of these compartments, remembering that N_i is identified with N_s, we obtain for M_s the number of particles in the shell s:

$$M_s = Q_s N_s = \alpha(4\pi/c^3 H)\epsilon_s^2 e^{-\epsilon_s/kT} d\epsilon_s \equiv F(\epsilon_s)d\epsilon_s \tag{38}$$

and evaluate α by the same procedure as before. The result is:

$$F(\epsilon) = \frac{N}{3k^3T^3} \frac{\epsilon^2}{e^{\epsilon/kT}}. \tag{39}$$

This is the smoothed-over distribution-in-energy predicted for the radiation-gas by the classical statistics, it being assumed that the momentum-space is to be divided into compartments of equal volume. Experiment however supplies a quite different distribution-in-energy, to wit:

$$F(\epsilon) = \left(\frac{8\pi V}{c^3 h^3}\right) \frac{\epsilon^2 d\epsilon}{e^{\epsilon/kT} - 1}. \tag{40}$$

$$U = \nu F$$

and the area between two ellipses is equal to h if the energy-difference between them is equal to $h\nu$.

It looks as if (39) might be the limiting form of (40)—as if the actual distribution-law might be obtained by avoiding the approximations whereby we came to the formula (39), as Planck's law of distribution for oscillators was obtained by refraining from approximation. Such however is not the case. True, the second factor in (39) is evidently the limiting form, for very high temperatures, of the second factor in (40). But the first factor in (40) contains nothing but the volume of the gas and some universal constants, while the first factor in (39) contains the temperature and an apparently disposable constant standing for the number of particles in the assemblage. The former is not the limit of the latter. It will be noted also that although I said that the volume of the elements of phase-space was to be set equal to h^3, this assumption in no wise enters into the function (39). Bose in fact found it necessary to upset the basis of the classical statistics, in order to arrive at (40) instead of (39).

The Bose Statistics

The momentum-space of the photons is to be divided as heretofore into equal compartments, and various distributions of the particles among these are to be compared, in order that we may elect one of them as "the most probable" and make a picture of the entropy of the assemblage. But the manner of defining a distribution, the manner of "counting the ways" in which it may be realized and computing its "probability," is to be changed, and changed in a most thoroughgoing and fundamental way.

Start with any distribution of the particles, defined as heretofore: defined that is, by saying that there are N_0 of the particles in the compartment 0, N_1 in the compartment 1, and in general N_i in the compartment i.

Count the number of compartments containing no particle; call it Z_0. Count the number of compartments containing one particle apiece; call it Z_1. In general, let Z_i stand for the number of compartments containing i particles apiece. Put down the values of all the numbers Z_i.

Now change the terminology. Elect some neutral word, "arrangement" say, to denote what we have heretofore denoted as a "distribution," and use the latter word in the following new sense: a *distribution* shall henceforth be described by stating the values of the numbers $Z_0, Z_1, Z_2, \cdots, Z_i, \cdots$ and the total energy of the assemblage.[7]

This means that each distribution in the new sense comprises a number of distinct distributions in the old sense. This we shall regard as the number of different ways in which the distribution in the new sense may be realized. Going over entirely to the new terminology: we shall now identify the probability W^* of a distribution with the number of arrangements which are included in it. Previously we identified the probability W of an arrangement with the number of permutations included in it, according to Eq. (7). This

[7] It is of course confusing thus to change the meanings of words, but in the long run less confusing (I think) than to use some other word than *distribution* for the concept always called by that name in the new statistics.

we now must forget; we must proceed as if the probability of each arrangement were the same.

The number W^* is now to be evaluated. In doing this we must remember that we have to count, not the total number of arrangements yielding the prescribed set of values of the quantities Z_i, but the portion of these which give the prescribed value to the total energy of the assemblage.

As before, we superpose upon the partitioning of the momentum-space into small compartments of equal volume H, another partitioning into spherical shells each of which is sufficiently large to contain many of the compartments, yet sufficiently small so that the same value of ϵ may be assigned to all the compartments within it. The final result is thus to be a "smoothed-over" formula. It is rather singular that whereas Planck introduced the quantum into physics by avoiding the smoothing-over which had been customary in the classical statistics, the quantum-formula for radiation is now derived by a method in which it is accepted.

Consider then any shell at random, say the "shell s." Denote by Q_s or by Z the total number of compartments in it; by Z_{is} the number of these compartments which contain i particles apiece; by M_s the number of particles in the shell. According to the scheme now being tried out, the number of ways of attaining the particular distribution characterized by the numbers Z_{is} is given by the formula:

$$W_s^* = \frac{Q_s!}{Z_{0s}!Z_{1s}!Z_{2s}!\cdots} \qquad Q_s = \sum_i Z_{is} \tag{41}$$

As the energy-values for all the compartments in the shell are (by hypothesis) approximately the same, these various ways of attaining the distribution Z_{is} all correspond to approximately the same total energy, as well as the same total number of compartments and the same total number of particles.

Suppose this process repeated for every one of the shells s. The total number of ways of attaining the actual distribution, compatible with the conditions of constancy of total energy, total number of particles and total number of compartments, is then the product of all the quantities W_s^*: call it W^*:

$$W^* = \prod_s W_s^*. \tag{42}$$

As before, and with the same end in view, we form the expression for $\log W^*$, and employ Stirling's formula (assuming thus in effect that none of the quantities Z_{is} is smaller than ten or so):

$$\log W^* = \sum_s \log W_s^* = \sum_s \left[Q_s \log Q_s - \sum_i Z_{is} \log Z_{is} \right]. \tag{43}$$

We now take the very great step of *identifying* not $\log W$ of equation (10), but $\log W^*$, *multiplied by a constant k, with the entropy of the assemblage.*

$$S = k \log W^*. \tag{44}$$

Then, when the numbers Z_{is} are changed by small amounts δZ_{is}, the ensuing change δE in the total energy E of the assemblage must be linked to the ensuing change in $(k \log W^*)$ by the equation:

$$\delta S = \delta(k \log W^*) = \delta E/T. \tag{14}$$

The first variation of $(k \log W^*)$ is given thus:

$$\delta(k \log W^*) = -k \sum_s \sum_i (1+\log Z_{is})\delta Z_{is}. \tag{45}$$

Let us try the distribution:

$$Z_{is} = \alpha_s e^{-i\epsilon_s/kT}. \tag{46}$$

Substituting this expression into (54), we get:

$$\begin{aligned}\delta(k \log W^*) &= -k \sum_s \sum_i (1+\log \alpha_s - i\epsilon_s/kT)\delta Z_{is} \\ &= -k \sum_s (1+\log \alpha_s)\sum_i \delta Z_{is} + \frac{1}{T}\sum_s \epsilon_s \sum_i i\delta Z_{is} \\ &= \delta E/T\end{aligned} \tag{47}$$

for the summation $\sum_i \delta Z_{is}$ vanishes because the number of compartments in each shell is invariable, while the quantity $\epsilon_s \sum i Z_{is}$ is equal to the total energy of the particles in the shell s.

The new statistics, in proposing a new conception of entropy as embodied in equation (44), therefore leads to a new distribution for thermal equilibrium. This distribution is expressed by equation (46) in the new-fashioned way, by stating the number of compartments in each shell which contain each of the permissible quotas of particles. We must translate it into a distribution-function-in-energy such as we used to express the results of the old statistics.

Before undertaking this translation, we compute the values of the constants α_s by summing the numbers Z_{is} over all values of i for each shell separately, and equating the sum to the total number Q_s of compartments in the shell. We obtain:

$$Q_s = \sum_i Z_{is} = \alpha_s \sum_i e^{-i\epsilon_s/kT} = \alpha_s (1-e^{-\epsilon_s/kT})^{-1}. \tag{48}$$

On substituting these values of α_s into (46) we get something which begins to look familiar.

Next for the number M_s of the particles in the shell s, we compute:

$$\begin{aligned}M_s &= \sum_i i Z_{is} = \alpha_s e^{-\epsilon_s/kT}(1-e^{-\epsilon_s/kT})^{-2} \\ &= Q_s \frac{1}{e^{\epsilon_s/kT}-1}\end{aligned} \tag{49}$$

which begins to look very familiar indeed.

Now for Q_s, the number of compartments in the shell s, we put the value already stated in equation (37), derived from the assumption that the compartments are all of the same volume H:

$$M_s = \frac{4\pi}{c^3 H} \frac{\epsilon_s^2}{e^{\epsilon_s / kT} - 1} d\epsilon_s \equiv F(\epsilon_s). \tag{50}$$

Here the function $F(\epsilon_s)$ is the "smoothed-over" distribution-in-energy in which the new statistics culminates. Unlike those which we earlier derived from the old statistics, it involves the volume of the elementary compartments directly. Whereas from the old statistics we obtained formulae involving the quantum only by avoiding the approximation, here we obtain a quantum-formula even when we admit the approximation——a contrast on which, I think, it is worth while to insist.

Let us then, in preparing for the final assumption, accept the principle generalized from Planck's assumption about oscillators: let the elementary cell of phase-space be given the volume h^3. This is not yet an assumption about the compartment of momentum-space. Supplement it, then, by supposing that the compartment of phase-space h^3 is the product of the compartment H of momentum-space and the entire volume V occupied by the radiation-gas. Then:

$$H = h^3 / V. \tag{51}$$

This assumption—let me remark in passing—takes a very elegant form if we replace the compartments by the permitted energy-values, and then the corpuscler picture by the wave picture; for then we have a series of permitted wavelengths, which are precisely those which can form stationary waves in a cube of volume V.
So the new statistics leads to the distribution-law:

$$F(\epsilon) d\epsilon = \frac{4\pi V}{c^3 h^3} \frac{\epsilon^2}{e^{\epsilon / kT} - 1} d\epsilon. \tag{52}$$

Dividing out the factor V, we get the number of particles per unit volume having energy-values between ϵ and $\epsilon + d\epsilon$, in an assemblage having the most probable distribution at the temperature T. Multiplying this by ϵ, we get the total energy per unit volume in the possession of such particles. Identifying these particles with photons, we observe that they have wave-lengths between ch/ϵ and $ch/(\epsilon + d\epsilon)$, frequencies between ϵ/h and $(\epsilon + d\epsilon)/h$. Transforming then from the variable ϵ to the new variables λ and ν, we obtain distribution-functions which give the *density of radiant energy* as functions of wave-length and frequency. It turns out that these agree absolutely with the observed distributions, except that they lack a factor 2. This factor is at once imported, and is ascribed to the fact that light is polarizable. So we arrive at the black body radiation-formula:

$$\rho(\nu) d\nu = \frac{8\pi h}{c^3} \frac{\nu^3 d\nu}{e^{h\nu / kT} - 1} \tag{53}$$

and the new statistics is justified by its success.

It will be observed that the new statistics leads to a precise value for the number of photons per unit volume, at any prescribed temperature; whereas the old statistics led to nothing of the sort, but to a formula which contained the number of atoms per unit volume as a disposable constant. This corresponds to a profound physical difference between radiation-gas and material gases. When I state the temperature and the volume of a box containing helium, I am not giving data enough to fix the quantity of helium inside the box; on the contrary, the quantity and the density of the helium in the box can be varied *ad libitum* while the temperature and the volume are held constant. But when I state the temperature and the volume of an enclosure containing radiation, I am giving data sufficient to fix the amount of radiant energy and the number of quanta in the enclosure absolutely. This is a fact of experience, and the new statistics is evidently in accord with it. But if one were tempted to try out the new statistics upon a material gas, would there be any way of avoiding the inadmissible conclusion that the number of atoms in such a gas is also absolutely fixed by temperature and volume?

There is such a way. One might replace the distribution proposed in equation (46) by a more general one involving a disposable constant B, as follows:

$$Z_{is}=\alpha_s e^{-iB-i\epsilon_s/kT}. \tag{54}$$

On substituting this into the expression for (k log W) we get instead of (47) the equation:

$$\delta(k \log W^*) = -k \sum_s (1+\log \alpha_s) \sum_i \delta Z_{is} + kB \sum_s \sum_i i\delta Z_{is} + \frac{1}{T} \sum_s \epsilon_s \sum_i i\delta Z_{is} \tag{55}$$

The right-hand member must as before reduce to $\delta E/T$ if the distribution (54) is acceptable; and this it will do, provided that not only $\sum_i \delta Z_{is}$ but also $\sum_s \sum_i i\delta Z_{is}$ is zero. Now the second of these quantities is zero for all variations in which the total number of particles remains the same. The distribution (54) enjoys a greater entropy than any other which is compatible with the same total energy and the same total number of particles. The distribution (46) was still more exalted; it enjoyed a greater entropy than any other compatible with the same total energy, even including those for which the total number of particles was somewhat different. But the distribution (54) is sufficiently distinguished to be qualified as the most probable distribution for a material gas. It seems rather singular that the distribution (46) is required for radiation-gas. Here is evidently one of the deep differences between matter and radiation.

Following the same routine as before, we arive at the following expression for the number of particles in the shell s:

$$M_s = Q_s \frac{1}{e^{B+\epsilon_s/kT}-1} \tag{56}$$

and in dealing with radiation-gas we have put $B=0$ and have taken the value of Q_s from equation (37). If in dealing with a material gas we take the value of Q_s from equation (25) instead and put $H=h^3/V$ we obtain:

$$M_s=\frac{2V}{h^3(\pi)^{1/2}}(2\pi m)^{3/2}\frac{(\epsilon_s)^{1/2}d\epsilon_s}{e^{B+\epsilon_s/kT}-1}\equiv F(\epsilon_s)d\epsilon_s \tag{57}$$

and now it is obvious that we must evaluate B in terms of the total number of particles N by the already so familiar way of integrating $F(\epsilon_s)$ over the entire energy-range from 0 to ∞ and setting the integral equal to N.

Einstein proposed this as an alternative to the Maxwell-Boltzmann law derived from the classical statistics. It is not easy to decide which of the two is supported by experiment, as with increasing temperature the formula (66) becomes more and more nearly like the classical one, and it turns out that throughout the convenient ranges of temperature and pressure the two are indistinguishable. It would be very valuable to determine between the two as then we should know which of the two ways of defining a distribution and estimating the probability thereof, which of the two pictures of entropy, is the proper one for a material gas. The reader may have remarked that if one were to apply Bose's method to the problem of determining the most probable distribution of particles in ordinary space, one would reach a result at variance with that of the classical statistics, and therefore at variance with the facts. One must deal altogether with the six dimensional phase-space, to be perfectly consistent. This is to be regretted.

The Fermi Statistics.

The statistics invented by Fermi, and later independently by Dirac, involves the same fundamental assumptions as that of Bose——the same manner of counting the ways in which a distribution may be realized, of defining its probability, of picturing its entropy. But there is an additional assumption, of the nature of a limitation: it is postulated, that a compartment may contain not more than some specific maximum number of particles. In particular for a gas towhich no external field is applied, it is postulated that each compartment must either be empty, or else contain one particle only.

The "exclusion-principle of Pauli" gave the hint from which the Fermi-Dirac theory sprang. This principle may be paraphrased as follows. In Bohr's "atomic theory of the atom" the electrons belonging to an atom are forbidden to revolve in any except certain specific orbits, set apart from the rest as the "permitted" orbits, and labelled by specific "quantum-numbers" In later versions of the theory the "permitted orbits" are less conspicuous, the "permitted quantum-numbers" more so; but the picture is acceptable at all events as a beginning. Upon this prohibition, then, Pauli superposed another; not more than one electron is allowed in each orbit or to each set of quantum-numbers. Perhaps it would be better to say "not more than some definite small number of electrons . . . . " instead of "not more than one." The affinity of this to one of Fermi's assumptions will soon be manifest.

It would take much too long to give an idea of the successes of the Pauli principle; they are however so great as to increase the inherent plausibility of Fermi's idea very much—or perhaps I should say, so great as to render the idea plausible, which otherwise it might not seem.

The reasoning follows exactly the same course as when we were deriving the distribution-law (56), except that all the summations over the variable i are now summations of two terms only, the term for $i=0$ and the term for $i=1$. For each of the shells there are only two numbers Z_{is} required to describe the distribution: viz. Z_{0s} the number of empty compartments and Z_{1s} the number of compartments containing one particle apiece. We try the distribution (54):

$$Z_{0s}=\alpha_s\ ; \qquad Z_{1s}=\alpha_s e^{-B-\epsilon_s/kT} \tag{58}$$

and easily find that it is the distribution of maximum probability, by comparison with all the others compatible with the same total number of particles and the same total energy. We arrive then at the following expression for the number of particles in the shell s:

$$M_s=Q_s\frac{1}{e^{B+\epsilon_s/kT}+1}\cdot \tag{59}$$

There is no point in putting for Q_s the value appropriate to radiation-gas, since the Bose formula has already proved adequate for that case. Fermi put the value appropriate to material gases, and obtained:

$$M_s=\frac{2V}{h^3(\pi)^{1/2}}(2\pi m)^{3/2}\frac{(\epsilon_s)^{1/2}d\epsilon_s}{e^{B+\epsilon_s/kT}+1}=F(\epsilon_s)d\epsilon_s. \tag{60}$$

This formula is the point of departure for the theory of the electron-gas in metals revived and remodelled by Pauli and Sommerfeld, to the experimental tests of which most of the rest of this article will be devoted.

Application of the Fermi Statistics to the Electrons in Metals

We are asked to conceive of a piece of metal as a region populated with "free" electrons, and surrounded by a wall; the electrons being distributed according to the formula of Fermi.

The Fermi distribution-function involves the total number N of the electrons, which is a disposable constant. It also involves the volume V which this assemblage of N particles pervades. For this we set the volume of the piece of metal—a decision which is tantamount to ignoring the atoms, to supposing the metal a vacuum inhabited by free electrons only. So remarkable an assumption, even though it be made only in approximation, requires some excuse. Its strangeness may be mitigated by recalling, first, that slow electrons may go through atoms (at least through certain kinds of atoms) as imperturbably as if the atoms were not there; and second, that a wave-train may go without being scattered at all through a crowd of particles individually

quite able to scatter it, provided that the particles are arranged in a regular lattice having a spacing smaller than the wave-length of the waves. The speeds attributed to electrons in metals are so low and their wave-lengths are so great, that perhaps they do behave in such a way.

The "wall" is the agency which prevents the electrons from escaping; it is commonly imagined as a sharp and sudden gradation of potential at the with a velocity of which the component normal to the bounding surface may be denoted by u, is supposed to be driven back into the body of the metal if the corresponding "component of kinetic energy" $\frac{1}{2}mu^2$ is less than a certain constant W_a; while if $\frac{1}{2}mu^2 > W_a$ the electron escapes, but with its kinetic energy diminished by W_a. According to newer ideas electrons may sometimes escape even when their values of $\frac{1}{2}mu^2$ are smaller than W_a, and may sometimes fail to escape in the contrary case; but the earlier and simpler conception remains approximately valid, and I will abide by it for a time. The constant W_a may be named the *work-function*.

Like the constant N, the work-function figures as a disposable constant in the theory. It is an ambition of physicists to explain as many as possible of the differences between different metals, by varying only the values of these two constants. Later we shall find it necessary to introduce others, beginning with the one which in the older theories appeared as the mean free path of the electrons; but there are several results of value which can be obtained with no other but these two.

I repeat now from (60) the Fermi formula for the distribution-in-energy of an assemblage of N particles in volume V at temperature T, with two changes made to bring the notation into harmony with that of Sommerfeld:

$$F(\epsilon) = G\frac{2}{(\pi)^{1/2}}\frac{V}{h^3}(2\pi m)^{3/2}(\epsilon)^{1/2}\frac{1}{A^{-1}e^{\epsilon/kT}+1}. \tag{61}$$

Here the symbol $1/A$ replaces e^B, and a factor G to which we shall assign the value 2 is introduced for a reason which will be mentioned later. The corresponding distribution-function in the coordinates and momenta is this:

$$f(x, y, z, p_x, p_y, p_z) = \frac{G}{h^3}\frac{1}{A^{-1}e^{\epsilon/kT}+1}. \tag{62}$$

The first step now is the same as in the classical statistics: to determine the constant A in terms of N by integrating $F(\epsilon)$ over the whole range of energy-values from 0 to ∞, and equating the integral to N:

$$\int_0^\infty F(\epsilon)d\epsilon = N. \tag{63}$$

This was an easy step in the classical statistics, but here it is very hard. The integral of $F(\epsilon)$ is not one of the common well-known functions to be found in mathematical tables, nor a combination of such; and we do not get a simple equation to be solved for A in terms of N. Sommerfeld indeed found it necessary to compromise by deducing two series-expansions for

the integral, one being available for values of A smaller than unity, the other for the opposite extreme. By a stroke of luck which seems almost too good to be true, the first one or two terms of one or the other of these expansions form an approximation amply good enough for all the cases where as yet theory and experiment can be compared.

I consider first the approximation which is of *no* importance in the theory of electrons in metals—the one for values of A so very small that the second term in the denominator of $F(\epsilon)$ is negligible by comparison with the first. Then the distribution-law approaches that of Maxwell and Boltzmann, and of necessity the constant A must possess the value which in the limit makes $F(\epsilon)$ identical with the classical expression written in equation (27): to wit, the value:

$$A = \frac{Nh^3}{GV}(2\pi mkT)^{-3/2}. \tag{64}$$

It would however be a great error to suppose that this value of A can be substituted into the function F under all circumstances. This value is acceptable only if A is very small relatively to unity, which is to say, if the quantity to which A is here equated is very small. So the question arises: in any physical case, is the combination on the right-hand side of equation (64) a small fraction of unity, or is it not?

Now for any material gas under any conditions usual in the laboratory, A is indeed very small. The new statistics leads to a result indistinguishable from that of the old statistics. To discriminate between the two by experiments on material gases, one would have to work with temperatures so low and densities so high that the gases would probably either be liquefied already, or at least would be in a condition very different from that "ideal" state to which the statistics is tacitly supposed to apply. Perhaps though it is not impossible to make the test with helium or hydrogen.

At this point apparently Fermi stopped. But it occurred to Pauli that if the new statistics were applied to an electron-gas as dense as that which Riecke and Drude had supposed to pervade the interiors of metals, the deviations from the classical distribution would be much more pronounced. For, in the first place, the mass m of the individual electron is smaller by several orders of magnitude than the mass of the atoms or molecules of any material gas. And, in the second place, if the number N of free electrons in a piece of metal is as great as or greater than the number of atoms, then it is thousands of times as great as the number of particles in an equal volume of a material gas. Now the quantity equated to A in equation (64) contains N in the numerator and $m^{3/2}$ in the denominator, and for the hypothetical electron-gas within the metals it is no longer small. The expression (64) for A is then no longer acceptable.

While the statement just made about m is based on a fact of experience, the statement about N is not so firmly grounded. We have no direct knowledge of the number of free electrons in a given volume, say the number n

$(=N/V)$ in unit volume, of a metal. This as I said above is a disposable constant of the theory. One of the tests of the theory is whether one can obtain correct numerical values of half-a-dozen properties of a metal by choosing a single value of n for that metal. So long as the classical statistics was applied to the electron-gas, this was impossible. If the value of n was put as high as the number of atoms in unit volume, the predicted value of specific heat (and we may now add, the predicted value of susceptibility) turned out to be too large; if n was lowered sufficiently to avoid this particular discordance, other predictions were impaired. It was however the general impression, that one should put n equal to the number of atoms or a small multiple thereof. I suspect that this decision was largely due to a feeling that since the free electrons are detached from atoms, and since all the atoms are alike, any atom should supply as many free electrons as any other. However that may be, it was natural though not inevitable for Pauli and for Sommerfeld to link the Fermi statistics with the postulate that there are as many free electrons as there are atoms, and test the combination of these two assumptions.

On putting for m the mass of the electron, for N/V the number of atoms per unit volume of any metal, for T any temperature from zero absolute up to several thousand degrees, and for G any small integer, one finds that the quantity equated to A in (64) is very large. Thus with $N/V = 5.9 \cdot 10^{22}$ (the number of atoms in a cc of silver), $T = 300°K$, $G = 2$, Sommerfeld computed:

$$(nh^3/G)(2\pi mkT)^{-3/2} = \text{about } 2400 \tag{65}$$

a result which invalidates equation (64).

We turn then to the other series-expansion of the integral $\int F(\epsilon)\, d\epsilon$, the one which Sommerfeld proved applicable for large values of A. The first two terms of this expansion are as follows:

$$\int_0^\infty F(\epsilon)d\epsilon = N = \frac{GV}{h^3}\,\frac{4\pi}{3}(2mkT\,\log A)^{3/2}\left(1+\frac{\pi^2}{8}(\log A)^{-2}+\cdots\right) \tag{66}$$

Taking the first term only of this expansion and putting the aforesaid values of N/V and T and solving for $\log A$, one finds a very large value indeed $(\log_e A = 325)$. Assuredly then we may use the first two terms of this expansion by themselves when we are dealing with the electron-gas in a metal, and indeed the first term will for some purposes be amply sufficient.

We have thus the following first—and second—approximation formulae for A in terms of n or N/V:

$$\begin{aligned} &2mkT\,\log A = h^2(3n/4\pi G)^{2/3} && \text{first approx.} \\ &2mkT\,\log A = h^2(3n/4\pi G)^{2/3}\left[1-\frac{(2\pi mkT)^2}{12h^4}\left(\frac{3n}{4\pi G}\right)^{-4/3}\right] && \text{second approx.} \end{aligned} \tag{67}$$

(the second approximation being computed by putting the first-approximation value of $\log A$ into the second term of the series expansion).

On substituting one or the other of these into the distribution-functions (61) and (62), we have the postulated distribution of the free electrons expressed to as high a degree of approximation as we require, with no disposable constant except n; and we are ready for the applications.

The Specific Heat

As it was the notorious difficulty with the specific heat which spoiled the old electron-gas theory in which the classical statistics was coupled with the assumption that there are as many free electrons as atoms, let us first of all find out whether this difficulty remains.

Any distribution-law for an assemblage leads immediately to a formula for the total energy E thereof as function of the temperature, which is:

$$E = \int_0^\infty \epsilon F(\epsilon) d\epsilon. \tag{68}$$

Putting the distribution-in-energy (28) derived from the classical statistics, we find:

$$E = \frac{3}{2} NkT \tag{69}$$

and putting the one just derived from the Fermi statistics with the first-approximation value of A, we find:

$$\begin{aligned} E &= E_0 + \tfrac{1}{2}\gamma V T^2 \\ E_0 &= \frac{2\pi}{5} \frac{VGh^2}{2m} \left(\frac{3n}{4\pi G} \right)^{2/3} \\ \gamma &= \frac{\pi}{3} n G \frac{(2\pi k)^2}{h^2} \left(\frac{3n}{4\pi G} \right)^{1/3} \end{aligned} \tag{70}$$

two exceedingly different formulae.

The derivatives dE/dT of these expressions are the formulae supplied by the two statistics for the specific heat of the electron-gas. The classical theory predicts for the specific heat a constant value, while the Fermi statistics makes it proportional to the temperature—being thus in harmony with Nernst's heat theorem, while the other is not— and gives it even at room temperatures but a small fraction of the classical value.

Experimentally the specific heat of the electron-gas cannot be measured separately from that of the lattice of atoms, which constitutes the metal—an admission which seems to condemn as vain all hope of testing these formulae. Nevertheless one can conclude with fair certainty that the classical expression is inadmissible; for the specific heat of an ordinary metal agrees so well with the value attributed by statistical theories both old and new *to the atoms alone*, that there is simply none left over for the electrons—no such great excess, that is to say, as the amount $3nk/2$ which the classical theory requires. The device of reducing n to so low a value that $3nk/2$

would be inappreciable makes trouble in other directions, as I have intimated. But with the new statistics the theoretical value γVT is inappreciable even when n is made as great as the number of the atoms and T as great as many hundreds of degrees. Probably no one who did not often lament the defeat of the old and so very desirable electron-gas theory by that hard fact about the specific heat will ever quite realize the rejoicing caused by this victory of the new, which by this achievement succeeded *a se faire pardonner* many deficiencies in other fields.

Features of the Fermi Distribution.

I will now mention some of the features of the Fermi distribution which has thus justified itself by passing its first test.

The most startling of these may be inferred from the distribution-function (62) or (61), by inserting the first-approximation formula for A presented in equation (67), and a new symbol W_i:

$$f=\frac{G}{h^3}\,\frac{1}{e^{(\epsilon-W_i)/kT}+1}\,; \qquad W_i=\frac{h^2}{2m}\left(\frac{3n}{4\pi G}\right)^{2/3}. \tag{71}$$

At the absolute zero the exponential term is either infinity or zero, according as the variable ϵ is greater or less than W_i. Therefore the density of the electrons in phase-space is constant and equal to G/h^3 for all energy-values less than W_i, zero for all values of energy greater than W_i.

This striking result can easily be deduced from Fermi's basic assumption, without any statistics at all. Absolute zero is by definition the temperature of the state, being in which the assemblage can give away no energy whatever. If not more than one electron may occupy any compartment of the phase-space, absolute zero is attained when there is an electron in every compartment from the origin outwards to a sphere which is centered at the origin, and which has just the volume needful to contain as many compartments as there are electrons. The number of electrons in unit volume of the phase-space is and remains equal to the number of compartments in unit volume, *i.e.* to the reciprocal of the volume of the elementary compartment, from the origin outward to this sphere; there it suddenly sinks to zero, and so continues. Cooling-down of an assemblage is settling-down of the particles into this the most condensed of all permissible arrangements; it is like crystallization upon a lattice, only the lattice is in the phase-space.

The foregoing statements may all be repeated, with the words *phase-space* replaced by *momentum-space*. In the momentum-space, a sphere of radius p, consequently of volume $4\pi p^3/3$, contains $(4\pi p^3/3)/(h^3/V)$ of the elementary compartments. If we set this number equal to the total number of electrons N, and solve the resulting equation for p, we get the radius of the sphere which would just contain all the electrons if there were one in each compartment. If we set the volume of the sphere equal to N/G and solve for p, we get the radius p_m of the sphere which would just contain all the electrons if there were G of them in each compartment. But this is the

maximum value of the momentum of the electrons, it is the momentum of the fastest of the electrons. The corresponding speed v_m of the fastest of the electrons is p_m/m, therefore given by the expression:

$$v_m = \frac{h}{m}\left(\frac{3n}{4\pi}\right)^{1/3} \tag{71a}$$

and the corresponding kinetic energy $\frac{1}{2}mv_m^2$ is the same as W_i.

It is expedient to set down for future reference the *mean values* of speed v and of several integer powers of v, for a gas distributed according to the Fermi law at the absolute zero. The general formula for the mean value of any power v^s of v is this:

$$\overline{v^s} = \frac{1}{n}\int v^s f(v)dv = \frac{1}{n}(4\pi G/m^3h^3)\int_0^{v_m} v^{s+2}dv$$

and in particular:

$$\overline{v^{-1}} = 3/2v_m \; ; \qquad \bar{v} = 3v_m/4 \; ; \qquad \overline{v^2} = 3v_m^2/5 \; ; \qquad \overline{v^3} = \tfrac{1}{2}v_m^3 \tag{72a}$$

The corresponding values for the Maxwell distribution are these:

$$\overline{v^{-1}} = 2(m/2\pi kT)^{1/2} \; ; \bar{v} = 2(2kT/\pi m)^{1/2} \; ; \overline{v^2} = (2kT/m) \; ; \overline{v^3} = 4\pi(2kT/\pi m)^{3/2} \tag{72b}$$

Plotted as functions of ϵ, the distribution-function f in the coordinates and momenta starts out as an horizontal straight line at a distance G/h^3

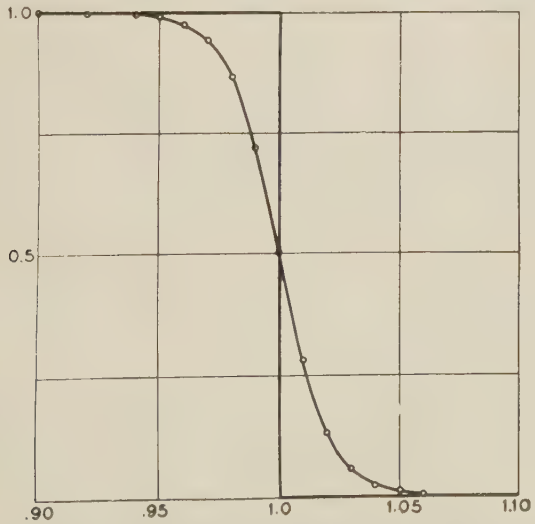


Fig. 1. Graphs of the Fermi distribution-function f plotted against ϵ/W_i asindependent variable, for an electron-gas having $6.5 \cdot 10^{22}$ particles per cc, at temperatures zero (rectilinear curve) and 1500°K, (rounded curve). The value of W_i is 6 equivalent volts.

from the axis of abscissae, while the distribution-function F in the energy starts out as a concave-upward parabolic arc; these continue as far as the

abscissa $\epsilon = W_i$, and from then on the curves coincide with the axis of abscissae.

The foregoing statements are valid for absolute zero; what happens as the temperature rises? Sommerfeld has proved that the sharp angles in the distribution-curve are *very gradually and slowly* rounded off, the curve always traversing the midpoint of the vertical arc BC (Fig. 1). The far end of the curve sinks down to the axis of abscissae in the fashion of the Maxwell law. Even at room-temperature and even far above, however, the distribution departs so little from the absolute-zero form that many phenomena may be interpreted in a qualitative way, simply by imagining the absolute-zero distribution—the completely degenerate distribution, it is called—to persist all through the observable range of temperatures. Indeed, in calculating electrical resistance and certain other properties of metals, one may use the mean values of the various powers of v which are tabulated in (72a). There are however other properties of metals, thermal conductivity for instance, for the estimation of which it is not sufficient to assume that the mean values of the powers of v are always the same as at absolute zero, and one must derive more nearly approximate values for them; for these however I refer the reader to Sommerfeld.

In practice, the values of W_i are rather astonishingly great; no less, for example, than 5.6 equivalent volts for silver, 5.7 for tungsten, 6.0 for platinum. Obviously they depend on the compactness of the lattice, being greater the more closely-packed the atoms are. In potassium and sodium the atoms are relatively widely spaced, and the corresponding values of W_i are about 2.1 and 3.2 in equivalent volts.

The contrast between this and the classical situation is evidently enormous. Where formerly we were asked to think of the electrons in a metal at usual temperatures as being distributed Maxwell-wise about a very modest mean energy, say about 0.02 of an equivalent volt, we are now invited to conceive them as distributed all through a range of energies extending from zero up to as much as half a dozen equivalent volts, and more abundantly the nearer one approaches to the top of this range, abruptly though the distribution ceases when the very top is reached. This is "zero-point energy" with a vengeance!

The pressure of the electron-gas is related to the energy-per-unit-volume by the equation valid also in the classical theory:

$$p = \frac{2}{3}\frac{E}{V}$$

and therefore varies like the total energy—starting from a value absurdly high at the absolute zero (hundreds of thousands of atmospheres) and increasing therefrom very slowly at first, though according to a T^2 law, as the temperature rises. I do not know of any manometer for measuring internal electron-pressures, but if anybody should invent one he had better make it strong.

There is manifest ground for doubting these remarkable proposals: thermionic data seem to show that the work-function which opposes the egress of the electrons from a metal is itself less than half-a-dozen volts (in the usual measure), for some metals less than two—what then keeps these fast electrons confined within the metal? It turns out, however, that in augmenting the *vis viva* of the electrons the new theory also raises the top of the wall which they must overleap. Here indeed we meet with the first of the new experiments which tend to confirm the new theory.

Thermionic Emission.

The simplest theory of the thermionic current is, that it consists of all the electrons belonging to the interior electron-gas which fly against the boundary-surface of the metal with velocities such that the component u thereof perpendicular to the boundary-surface is great enough to make the "energy-component" $\frac{1}{2}mu^2$ greater than a constant W_a—the said constant being interpreted as the work-function or the retarding potential-drop at the edge of the metal. The thermionic electrons are those which swim up to the surface with an outward-bound velocity-component so large, that by means of the kinetic energy of their outward motion they can climb over the wall.

Evidently any thermionic emission must distort the distribution of the electron-gas inside the metal, as it is an unbalanced outflow of electrons. The situation in which the efflux is balanced by a corresponding influx from an electron-gas outside the metal is much regarded in thermodynamic theory, but one cannot measure currents in that situation any more than one can measure heat-flow between two bodies at equal temperature. In assuming, then, that the distribution of the internal electrons is that of Fermi or that of Maxwell, we shall probably be invalidating our conclusions except for the limiting case of an infinitesimal emission. It seems probable, however, that with the thermionic currents of practice the approximation is good enough.

The simplest theory of the thermionic current, then, consists entirely of the equation:

$$i=e\left(\frac{m}{h}\right)^3 G\int_{u=u_0}^{\infty}du\int_{-\infty}^{\infty}dv\int_{-\infty}^{\infty}dw\cdot u\frac{1}{A^{-1}e^{\epsilon/kT}+1} \tag{73}$$

This is a restatement of the first sentence of this section, *plus* the assertion that even when electrons are leaking out through the wall of the metal the distribution within remains practically that of Fermi. The factor e stands for the electron-charge; the symbol i thus for the thermionic current-density in electrostatic units. The factor m^3 enters because, in conformity with usage, I have translated from the momenta into the velocity-components u, v, w as independent variables. The quantities u_0 and W_a are related by the equation:

$$W_a=\tfrac{1}{2}mu_0^2 \tag{74}$$

In the integrand we are of course to put $\frac{1}{2}m(u^2+v^2+w^2)$ for ϵ.

Setting for A the first-approximation value from (107) with the symbol W_i defined in (71), we obtain:

$$i = e\left(\frac{m}{h}\right)^3 G \int_{u_0}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u \frac{1}{e^{(\epsilon - W_i)/kT}+1} du dv dw \tag{75}$$

The integration is perfectly straightforward if the second term in the denominator may be neglected relatively to the first. This seems unnatural, for we have just been noticing that over part of the energy-range, from $\epsilon = W_i$ downwards, the second term is larger than the first. But if W_a is considerably larger than W_i—and by this I mean, if $(W_a - W_i)/kT$ is positive and considerably larger than unity—then the electrons which escape are those which belong to the extreme upper part of the energy-range, where the first term is much the larger. Writing then $u \cdot \exp[-(\epsilon - W_i)/kT]$ for the integrand, we integrate with ease by well-known formulae, and get:

$$i = \frac{2\pi emG}{h^3}(kT)^2 e^{-(W_a - W_i)/kT}. \tag{76}$$

The experimental test consists in plotting $(\log i - 2 \log T)$ against $1/T$; we should get a straight line provided that W_a does not vary with temperature. The experiments do lead to precisely this result. The slope of the line varies from metal to metal, and depends on the state of the metal surface; identifying it with $(W_a - W_i)/k$, one finds that W_a exceeds W_i by amounts ranging from one equivalent volt upward to five or six; so the approximation just mentioned is abundantly justified, even up to temperatures of incandescence.

The contrast with the predictions of the classical theory is peculiarly interesting. Assuming the Maxwell distribution for the interior electrons, one arrives easily (the reader can do it by substituting the value of A from (64) into (73)) at the formula:

$$i = \frac{en}{(2\pi m)^{1/2}}(kT)^{1/2} e^{-W_a/kT}. \tag{77}$$

On testing this formula by plotting $(\log i - \frac{1}{2}\log T)$ against $1/T$, it is found that the experiments yield lines as beautifully straight as those obtained by plotting $(\log i - 2 \log T)$. Indeed—as everyone knows who has dabbled in thermionics—a function of the type $\exp(-c/T)$ varies so exceedingly rapidly with $1/T$ that it makes no perceptible difference to the graph whether or not the function is multiplied by a constant or by any modest power of T. One cannot then use the graphs to distinguish between the theories, even if one could be sure that W_a is not a function of T. But the classical theory proposes that we identify the slope of the aforesaid line with W_a/k; and it has been the custom so to do.

Now *if the Fermi distribution-function is the right one, physicists have have been underestimating the work-function all along.* They have plotted experimental curves which agreed in shape with (76) and (77) and from

these they have evaluated the constant figuring in the exponent, a constant which they have denoted usually by $-b/k$; and then they have equated b to W_a, whereas if it is right to apply the new statistics they should have added W_i to b and then equated the sum to W_a. Nor is the alteration slight: for if there are as many free electrons in the metal as there are atoms, then W_i is six volts or thereabouts, and the quantity to be added to the observed constant b is larger than b itself.

Is there then any other way of determining the work-function than out of this apparently ambiguous current-vs-temperature curve? If there is a direct and independent way, it may serve not only to decide between the two statistics, but also—if it favors the new by yielding a value for W_a greater than the thermionic b—to give an *experimental* value for $(W_a - b) = W_i$ and hence for the disposable constant n which is the one uncertain quantity in the theoretical formula for W_i. It may, that is to say, serve to determine the number of electrons per unit volume of the electron-gas.

Now it seems that the diffraction of electrons by crystals provides an independent and direct way of of determining the work-function. For in the phenomena in which negative electricity behaves as a wave-motion, the work-function figures in the index of refraction; and the index of refraction of a metal may be determined from the diffraction-patterns which it forms when irradiated with slow electrons. Ample data concerning one metal—nickel—have already been acquired by Davisson and Germer, and from these it transpires that the work-function is much in excess of the thermionic constant b—so much indeed, that the corresponding value of W_i implies that there are twice as many free electrons as atoms in the metal, or even more[8] The values deduced for the work-function from the refractive index vary however with the speed of the electrons; and it is evident that much remains to be understood.

The factor which multiples $T^2\exp[-(W_a - W_i)/kT]$ in the right-hand member of (76) involves universal constants only (supposing that G is such) and is therefore the same for all metals—a principle derived by Richardson from the first and second laws of thermodynamics twenty years ago, without any assumptions at all about the distribution of the electrons. Its actual value $2\pi k^2 meG/h^3$ differs only by the factor G from the value derived by Dushman, which is numerically equal—in the customary units—to 60.2 amperes per cm^2 per degree squared. There are several metals for which the experimental value of this quantity—commonly known as A, a symbol which in this article is monopolized by another meaning—agrees well with 60.2. One might infer that G must be unity, a choice which would demolish the theory of paramagnetism; but there is another recourse; one may suppose that half the electrons which come up to the bounding surface from within with energy sufficient to escape are nevertheless reflected. A factor $(1-r)$

[8] L. Rosenfeld, E. E. Witmer (*l. c. infra*). From Rupp they cite values of refractive index for six other metals (Al, Cr, Cu, Ag, Au, Pb) and compute values of n. In considering these, however, the reader should assess G. P. Thomson's criticism of Rupp's values.

$-r$ being called the coefficient of reflection, and being put equal to $\frac{1}{2}-$then enters into the formula, and balances out the factor 2 introduced by giving the preferred value to G. Moreover one may explain values of the constant still smaller than 60.2, or between 60.2 and 120.4, by adjusting r accordingly. But there are also recorded values enormously greater than 120; so evidently something remains to be understood.[9]

The methods of wave mechanics have been applied by Fowler and Nordheim to the problem of evaluating this coefficient of reflection. They have attained some notable results in the fields of thermionics, cold discharge, and photoelectric effect. These however are consequences not of the new statistics only, but of a combination of the new statistics with the new way of considering the transmission and reflection of electron-waves at surfaces. There is not space for me to deal with the latter, beyond indicating its point of departure and its chief results.

Thus far I have been speaking of a metal as an equipotential region surrounded by a surface at which there is a sharp potential-drop, and beyond which there is the equipotential region of outer space. Fowler and Nordheim however, like Schottky and others before them, conceive a metal as an equipotential region surrounded by a surface, beyond which lies a region in which there is a field (or at all events an image-field) the strength of which is a function of the distance from the surface. The quantity W_i appears as the integral of this field-strength, from the surface to infinity. Electrons of a given kinetic energy being supposed to fall against the surface from within, the fraction which fails to pass completely through the region of the field depends upon the kinetic energy of the electrons and upon the *shape* (not solely upon the integral) of the field-strength-vs-distance curve. The average value of this fraction for all the electrons of all speeds coming up to the surface from within —the average being taken with due regard to the relative proportions of the electrons of various speeds, that is to the distribution-in-velocity—is the coefficient r aforesaid. For certain simple shapes of the field-vs-distance curve it may be calculated. One thus arrives at the beginnings of a theory of the effects of surface-conditions on thermionic emission, and of the cold discharge.

The Cold Discharge

Suppose that a potential-difference is now applied between the metal and a neighboring electrode, such that near the metal the resulting field-strength is very great. Does it penetrate the region which I have just been describing? Assume that it does penetrate as far as the "surface" just defined. Then, everywhere beyond the surface, the actual field is the resultant of this "applied" and the previously-mentioned "intrinsic" field. The shape of the field-strength-vs-distance curve is thus changed, in a way which is calculable

[9] If the empirical equation is of the form $i = aT^n \exp(-b/T)$, then whatever may be the values of the constants a and b, one can always claim that it agrees with the foregoing theory provided one assumes that W_i or r or both vary with temperature in just the proper way. But, as Fowler puts it: "the variety of possible uncontrollable hypotheses (if such assumptions are to be admitted) becomes too large for profitable disscussion."

if we have postulated some particular original shape; and in certain simple cases it is possible to calculate the consequent change of the coefficient r, and therefore the electron-current—or the additional electron-current—which the applied field causes to emerge from the metal. This is the current known as the "cold discharge."

Nordheim adopted for the field-strength-vs-distance curve, in the absence of applied P. D., the shape which is most commonly proposed—the inverse-square curve, the law of the "image-force" which a charge in the vicinity of a conductor experiences because of its "electrical image" in the conductor. For the current i of the cold discharge he then derived this approximate formula:

$$i = c'SF^2 \exp\ (-c''/F) \tag{78}$$

in which F stands for the applied field, S for the surface-area of the metal exposed to it, c' and c'' for constants which can be calculated when W_i is known.

There are certain experiments (for an accont of which I refer to Nordheim's paper in the Physikalische Zeitschrift) which indicate that the actual relation between the current and the field-strength agrees in form with (78), but that the predicted values of c' and c'' are too large—too large by a factor of the order of ten in the latter case, by several orders of magnitude in the former. However it is possible to explain away these contradictions. One may assume, for instance, that the actual surface concerned in the discharge is a collection of small spots which altogether have but a small fraction of the total area of the metal surface; and that over these small spots, the field-strength is much higher than it would be if the metal were everywhere uniform and smooth. The ratio of the observed to the predicted value of c' then gives the fraction of the total surface which is covered by these "active" spots, and the ratio of the observed to the predicted value of c'' gives the reciprocal of the factor by which the field-strength must be multiplied.

This explanation has the disadvantage of being not only plausible but much too easy. So long as there is not any independent evidence about the area of the effective spots or the field-strength prevailing over them, the theory simply delivers an equation with two disposable constants, which is not very valuable for testing the underlying assumptions. It appears from the data examined by Nordheim, however, that the ratio of the predicted to the observed value of c'' is always between 10 and 20, and the ratio of the observed to the predicted value of c' is always about 10^{-10} (at least for tungsten). This uniformity of the two quantities which figure in the theory as the disposable constants may be taken as a confirmation of some weight.

Photoelectric Effect.

According to the former theory, the elementary process of the photo-electric effect runs thus: a quantum dives into a metal, and gives its whole energy (say E_0) to an electron initially at rest, which then may escape from

the metal after suffering a reduction of kinetic energy equal at least to W_a and possibly more (more, that is to say, if the electron loses kinetic energy on its way to the surface). Even if we suppose that the electron originally belonged to an electron-gas conforming to the classical statistics, its initial energy would almost always be quite negligible compared to that which the quantum gives it. But if the electron-gas obeys the new statistics, it comprises electrons with energy-values ranging up to W_i. However if the new statistics is valid, then the reduction of kinetic energy at the boundary is also greater by W_i than we have hitherto supposed. The net result is, that by the new statistics as by the old we derive Einstein's equation for the maximum kinetic energy of the electrons expelled by quanta of frequency ν:

$$E_{max} = E_0 + W_i - W_a = h\nu + \text{const.} \tag{79}$$

only the additive constant is now $(W_i - W_a)$ instead of $(-W_a)$. This additive constant should as before agree with the thermionic constant b.

The new theory has one marked distinction, probably an advantage, over the old: it implies a sharply definite maximum kinetic energy—that is to say, the distribution-in-energy function of the escaping electrons should jump suddenly from zero to some definitely higher value at the energy-value $E = E_{\max}$ prescribed by (79); the slope of the curve representing this function should make an acute angle with the axis of E where they intersect at $E_{\max}$. The Maxwellian distribution predicted by the classical statistics for the interior electrons suggests however that the curve in question should approach the axis asymptotically. The new statistics leads also to certain inferences about the shape of the curve for values of E less than $E_{\max}$. A great quantity of data bearing on this subject has been obtained by Ives and his collaborators; but the interpretation is made difficult by the presumption that some allowance must be made for the energy-losses suffered by electrons after they absorb quanta but before they reach the surface, and will require much study.

Paramagnetism of the Electron-Gas

The susceptibility of the electron-gas was calculated by Pauli even before the specific heat was evaluated by Sommerfeld, but as it involves an extra complication I have inverted the historical order.

The complication is due of course to that assumption which is made in order to explain why the electron-gas should be magnetic at all—the assumption that electrons are magnets. Perhaps I am too cautious in referring to it as an assumption, it being so well authenticated by the gyromagnetic effect and by the general usefulness of the "spinning electron" in the explanation of spectra. These phenomena impose a special value on the magnetic moment μ_0 of the electron, to wit, the value,

$$\mu_0 = eh/8\pi m_0 c \tag{80}$$

m_0 standing for the rest-mass of the electron. Further they require that when the electron is floating in a magnetic field, its moment (considered

as a vector) shall be either parallel or anti-parallel to the field. Denote by θ the angle between the moment of the electron and the magnetic field: then θ must be either 0 or π.[10] Now when a magnet of moment M is inclined at an angle θ to a magnetic field H, its "extra magnetic energy" is $-MH\cos\theta$.[11] In dealing with the electron-gas, then, we are in effect assuming that when a field H is applied to it the energy of every electron is either increased or decreased by the amount:

$$\Delta = ehH/8\pi m_0 c \tag{81}$$

As I mentioned earlier, the idea of compartments in the momentum space may be replaced by the idea of "permitted energy-values", at least in some situations. Let us for convenience return to the latter conception. Then we may say that when a magnetic field is applied to an electron-gas of which the particles are magnets, each of the permitted energy-values is split into two. To any previously-permitted value ϵ correspond a pair of new ones, $\epsilon+\Delta$ and $\epsilon-\Delta$; or let me say $\epsilon+m\Delta$, using m as a symbol which may have only the values $+1$ and -1.

Pauli assumed that the most probable distribution of the electrons among this doubled set of energy-values is to be determined by the new statistics, including Fermi's postulate so modified as to state that not more than one electron may possess any one of the permitted values.

Previously I used the symbol Z_{is} to denote the number of compartments or permitted energy-values which lie in the shell s and are occupied by i electrons apiece; and the symbol Q_s to denote the total number of permitted energy-values in the shell s. Now there are Q_s permitted values which are shifted upward by Δ from the original ones, and Q_s more which are shifted downward by Δ from the original ones. Let Z_{is+1} stand for the number in this upward-shifted group which are occupied by i electrons apiece, and Z_{is-1} for the corresponding number in the downward-shifted groups; Z_{ism} shall be the general symbol for the two.

Now consider the distribution:

$$Z_{ism} = \alpha_{sm} e^{-iB - i(\epsilon_s + m\Delta)/kT} \qquad \begin{matrix} m = +1, -1 \\ i = \ 0, \ \ 1 \end{matrix} \tag{82}$$

This answers the standard requirements for thermal equilibrium. For if we define W^* in Bose's way, and then say that the entropy is $k \log W^*$, we find that when the energy of the assemblage is varied by δE—the total number of cells and the total number of particles remaining constant—the first variation of the entropy is $\delta E/T$, as it should be. I leave it to the reader to prove this statement as corresponding statements for other distributions

[10] It comes to the same thing, and may on other grounds be preferable, to assign to μ_0 twice the value given in (80), and to θ the values 60° and 120°.

[11] The magnetic energy, or energy due to the "interaction between the magnet and the field" is put equal to zero when the magnet is transverse to the field, which is consistent with the picture that the field alters the energy of the magnet by speeding up or slowing down the revolving electricity. The formula here given is a first approximation.

(*e.g.* (47)) were earlier proved, and to determine the values of the quantities α_{sm}; on doing which, and substituting the results into this distribution, he should obtain

$$Z_{1s,+1}=\frac{Q_s}{e^{B+(\epsilon_s+\Delta)/kT}+1} \qquad Q_s=\frac{2}{(\pi)^{1/2}}\frac{V}{h^3}(2\pi m)^{3/2}(\epsilon_s)^{1/2}. \tag{83}$$

$$Z_{1s,-1}=\frac{Q_s}{e^{B+(\epsilon_s-\Delta)/kT}+1}$$

It is best, perhaps, to regard these formulae as the descriptions of two electron-gases, one composed entirely of magnets parallel to the field, the other of anti-parallel magnets; the actual electron-gas is a mixture of the two. The quantity $Z_{1s,-1}$, for example, is the number of parallel magnets having energy-values shifted downwards through Δ from the Q_s originally-permitted energy-values which lay in the shell s. True, these magnets are no longer themselves in the shell s, owing to the shift; their energies lie between $(\epsilon_s-\Delta)$ and $(\epsilon_s-\Delta+d\epsilon_s)$, not between ϵ_s and $\epsilon_s+d\epsilon_s$; but for ease of integration it is better to think of them as being associated with the original unshifted energy-values. The total number of magnets comprised in the parallel gas is then given by the equation:

$$N_1=\frac{2}{(\pi)^{1/2}}\frac{V}{h^3}(2\pi m)^{3/2}\int_0^\infty \frac{(\epsilon)^{1/2}d\epsilon}{e^{B+(\epsilon-\Delta)/kT}+1}. \tag{84}$$

For brevity denote by L the constant before the integral; and use the symbol ϕ for the function defined as follows:

$$\phi(u)=\int_{\epsilon=0}^\infty \frac{(\epsilon)^{1/2}d\epsilon}{e^{u+\epsilon/kT}+1}. \tag{85}$$

Then expanding N_1 as a power-series in the variable Δ/kT which is $\mu_0 H/kT$, we find:

$$N_1=L\left[\phi(B)-\frac{\mu_0 H}{kT}\phi'(B)+\text{terms of higher order in } H\right]. \tag{86}$$

Similarly one obtains, for the total number N_2 of electron-magnets in the "anti-parallel gas," a formula:

$$N_2=L\left[\phi(B)+\frac{\mu_0 H}{kT}\phi'(B)+\text{terms of higher order in } H\right]. \tag{87}$$

Now the total magnetic moment of the "parallel gas" is $N_1\mu_0$ in the same sense as the field, and the total magnetic moment of the "anti-parallel gas" is $N_2\mu_0$ in the sense opposed to the field; so that the net magnetic moment of the entire assemblage of electrons is $(N_1-N_2)\mu_0$. We will carry out the computations only for values of H so low that we may ignore all terms beyond the second in the expansions for N_1 and N_2. The net magnetic moment is then approximately proportional to H; its quotient by H, the

susceptibility χ of the electron-gas, is constant. It is a fact of experience that with nearly all paramagnetic substances the susceptibility *is* independent of H up to the highest attainable values of this variable. The limitation which we are here accepting will probably therefore not prove serious. Our approximative formula for χ is then as follows:

$$\chi=(N_1-N_2)\mu_0/H=-\frac{2L\mu_0^2}{kT}\phi'(B)=-\frac{N\mu_0^2}{kT}\frac{\phi'(B)}{\phi(B)} \tag{88}$$

and all that remains is to make the step made in every previous case—to determine the last remaining unspecified constant, B, in terms of the total number N of the particles of the assemblage.

This number N is the sum of N_1 and N_2. Ignoring the terms of higher order in H, we have:

$$N=2L\phi(B) \tag{89}$$

and this is substantially the equation which was used to determine Sommerfeld's constant A in terms of N; for e^B and $1/A$ are one and the same. To make this equation identical with (63), or rather to make (63) identical with this one, we must there put $G=2$, as we did—this is the reason for having introduced that factor G.

The procedure is then as follows: put $-B$ for $\log A$ in the righthand member of equation (66)—differentiate it with respect to B—insert into the derivative the value of B obtained by equating to N the right-hand member of (66), *i.e.*, the value given in (67)—and substitute into (88). The resulting value for χ is this:

$$\chi=12\left(\frac{\pi}{3}\right)^{2/3}\mu_0^2 n^{1/3} m_0 h^{-2}. \tag{90}$$

To pass now to the experiments: is it permissible to suppose that the susceptibility of any metal is due entirely to the electron-gas within it? This is the same sort of uncertainty as confuses the question of the specific heat. Here we have every reason to expect that the magnetization of an ordinary paramagnetic metal is a threefold effect, involving not only the orientation of the electrons but also the orientation of the atoms, and finally that alteration of the electron-orbits in the atoms which gives rise to diamagnetism. To disentangle these three contributions to the net magnetic moment seems almost beyond the powers of any theory. With the alkali metals, however there is strong evidence that the second may be absent. Spectroscopic data show quite definitely that the magnetic moment of the alkali-metal ion—the atom minus its valence electron—is zero. If every atom in an alkali metal has surrendered its valence electron to the electron-gas, then there will be no orientation of the ions by the magnetic field, and the number of electrons forming the electron-gas will be equal to the number of atoms. The values of χ computed with this last assumption should then be not less than the actual susceptibility; they may be somewhat

greater, because of the diamagnetic effect which is opposed in sign to the paramagnetic effect and therefore neutralizes it in part. Of this diamagnetic effect we can predict the order of magnitude, and we may expect that it will be greater, the higher the atomic number of the metal.

The value of χ proposed above is the value for absolute zero; for higher temperatures a closer approximation can be obtained by using two terms of the expansion in (66), instead of the first term only. It appears, however, that the alteration is slight. Like the average energy and the pressure, the susceptibility of the electron-gas should be very nearly the same at all temperatures from absolute zero up through room-temperature and far beyond. Now it is a fact that the susceptibility of the alkali-metals is independent of temperature—a fact so surprising, that the desire to explain it seems to have been Pauli's principal incentive in undertaking this research. For if the electron-gas were governed by the classical statistics, and the electrons were as many as the atoms, the susceptibility of a metal would increase as the temperature diminished and attain enormous values near the absolute zero.

When Pauli published the theory to which this section is devoted, the experimental data indicated that the susceptibilities of potassium and sodium were somewhat lower, those of rubidium and caesium markedly lower than the predicted values—divergences which might be charged to the diamagnetic effect or to faults in the theory. Recent Canadian work, coming out very much *a propos*, has improved the situation remarkably. This tabulation (taken from E. S. Bieler, to whose article I refer for the sources) shows the comparison:

	Na	K	Rb	Cs
Theoretical (Pauli)	0.66	0.52	0.49	0.45
Experimental:				
McLennan et al.	0.61	0.42	0.31	0.42
Lane	0.65	0.54		

(All numerical values to be multsplied by 10^{-6}

Singularly enough, the agreements are too good! one would expect the diamagnetic effect to be more considerable than the very slight discrepancies between the experimental and the theoretical values for sodium, potassium, and caesium. Perhaps further work on the theory of the diamagnetic effect would now be desirable.

Returning once more to the meaning of G, one sees that the placing of the value 2 for G in equation (61) and all of its descendants amounts to the making of the assumption that the electron-gas is really a mixture of two equally numerous and entirely independent assemblages of particles, each for itself obeying the Fermi statistics. This seems a rather odd idea, but inevitable.

Theory of Conduction.

The new theory of conduction developed by Houston and Bloch is based upon the wave-theory of negative electricity, in which the interior of

a metal is conceived to be filled not with darting corpuscles, but with stationary waves—as many distinct patterns of loops and nodes, it may be, as in the corpuscle-picture there are free electrons. It is not a consequence of the Fermi statistics alone, but of the Fermi statistics plus the wave-theory. Of course, if we come to decide that the Fermi statistics implies the wave-theory and *vice versa*, this warning will seem superfluous; but it is not superfluous, so long as the new statistics is used with reference to corpuscles. Now the corpuscle-picture of negative electricity is not only familiar, but seems likely to survive as the most convenient for describing most of the phenomena in which electrons figure. I will therefore express as much as possible of the new theory of conduction in the language of corpuscles, although eventually I shall be forced to make an assumption which will come to the same result as converting the corpuscles into waves.

To realize the things to be explained, conceive a slab of metal, having a thickness d measured along the x-axis; suppose a potential-difference V to exist between its faces, so that a field $E = V/d$ directed along the axis of x pervades it.

If the electrons in the metal moved perfectly freely, then any which was introduced without kinetic energy at the negative side of the slab would fall forthwith to the positive side, arriving there with the full kinetic energy eV and the full corresponding velocity of magnitude $(2eV/m)^{1/2}$ directed along the axis of x. Certainly nothing of the sort occurs. When a potential-gradient exists along a wire, for instance, heat is developed uniformly everywhere and there is nothing to suggest that the electrons are moving more rapidly at the positive than at the negative end.

We must then suppose that the free flight of the electron is interrupted at frequent intervals, and that at every interruption it loses the kinetic energy and the component of velocity up the potential gradient which it has acquired from the field since the last one previous. Or at least, the average loss of kinetic energy and of "drift-speed" at interruptions must be balanced by the average gain between interruptions.

In the corpuscle-theory these interruptions are pictured as actual impacts or collisions of the electrons with the atoms. Evidently, if we could assume that whenever an electron hits an atom it rebounds in some direction perfectly transverse to the field, then we should have a mechanism in which the drift-speed of the electron up the potential-gradient is annulled at every impact. This would be much too artificial. But if we think of both the electrons and the atoms as elastic spheres, the latter being so massive that they never budge when struck, the result is in effect the same. For then, the angle between the direction along which an electron approaches an atom and the direction along which it flies away after collision is *on the average* 90°. The rebound is as likely to be backward as forward; the rebounding sphere retains on the average no memory of its former direction of flight. This I will prove later.

There is a difficulty, which I must not leave unmentioned, although in this place I can do nothing to clear it away. In the development of these ideas

we shall in effect assume that at the end of each free path the electron loses not only the forward drift-speed but the whole of the kinetic energy which it acquired while traversing that free path under the influence of the field. But if it collides with infinitely massive spheres it does not lose kinetic energy at all. If it collides with spheres of the mass of an atom, it loses kinetic energy, but does not completely lose its drift-speed. The theory of this latter case has been developed by Compton and Hertz for use in the study of conduction in gases, and might be applied to the problem presented by metals, but probably fits them no better than does the other hypothesis.

With this elastic-sphere model, then, the average interval between impacts is the average interval during which the electron is piling up drift-speed, only to lose it all at the end of the interval and be forced to start afresh from scratch. Denote by t_0 the length of this average interval. Since the acceleration of the electron is eE/m, its drift-speed at the end of the period t_0 is (eEt_0/m), its average drift-speed is half as great. Now I must dispel the impression that the drift-speed is the whole of the speed which the electron has. On the contrary, the mean speed of the thermal agitation—let me call it $\bar{v}$—is immensely larger than the small contribution which any ordinary field (indeed, any not very extraordinary field) can impart to an electron over a distance comparable with the distance between atoms. The field must not be supposed to do more than bend very slightly the rectilinear paths of the electrons from impact to impact. This statement is true with the classical statistics, *a fortiori* with the new. Denote by l the average distance traversed by an electron between impacts. Then t_0 is $l/\bar{v}$, and the average drift-speed is $\frac{1}{2}(eEl/m\bar{v})$.[12] The corresponding current-density is the product of this by the number of electrons in unit volume multiplied by the charge of each. So, for the current-density produced by unit-field-strength, which is by definition the *conductivity* σ, we obtain the formula:

$$\sigma = \tfrac{1}{2}ne^2l/m\bar{v}. \tag{91}$$

The constant l, the *mean free path*, is the third disposable constant of the theory of electrons in metals.

I fear that the foregoing passage sounds very old-fashioned; but nevertheless it expresses the corpuscle-theory of conduction. The notion of elastic spheres is only accessory—an image which may or may not be the best to represent the central idea, the idea that the life-history of a corpuscle in a metal pervaded by a field is an alternation of gradual gain and sudden loss. The mean free path is the average distance of uninterrupted gain.

The common test of the formula (91) is the test by the temperature-variation. The result of this, incidentally, was regarded as quite as grave a demerit of the old electron-gas theory as the difficulty with the specific heat.

[12] It is the mean of the reciprocal of the speed, not the reciprocal of the mean speed, which should figure here; but with so rough a formula the distinction is scarcely worth making.

It is a fact of experience that the resistivity $\rho = 1/\sigma$ of any metal varies rapidly with temperature. For many metals it varies directly as T over quite a wide range; at low temperatures even more swiftly, not to speak of the strange phenomenon of supraconductivity. Now in equation (151) we have ρ set equal to a combination of two universal constants with three quantities $\bar{v}$, n, l between which last the responsibility for these great variations must be divided.

According to the classical statistics $\bar{v}$ is proportional to $T^{1/2}$. This is a variation in the right sense, but not fast enough. To make ρ vary as T we must then make nl vary as $1/T^{1/2}$. With the Fermi statistics the requirement is harder. The mean speed of thermal agitation is almost independent of temperature, and the burden of the whole responsibility for making ρ proportional to T must be loaded upon nl. The first step with the new statistics is a step backward.

Can we reasonably assume n to be the cause of the variation? If so, then it must diminish with rising temperature. It would seem reasonable enough for n to *increase* with rise of temperature, for presumably the free electrons arise from ionization of the atoms, and ionization is promoted by heat; but for n to *decrease* would seem very odd, notwithstanding Waterman's successes in accounting for some of the data by such a theory.

Part of the burden, then, must be cast on the mean free path—indeed the whole of it, if we adopt the new statistics so that $\bar{v}$ is held constant, and suppose in addition that n does not vary with temperature. But the elastic-sphere conception cannot stand the strain. It gives for the mean free path a value independent of temperature,[13] except insofar as the metal dilates with increase of heat. This is pretty nearly checkmate.

If however we might suppose that an electron may sometimes go clear through an atom without being reflected or deflected, and that the chance of such a piercing is relatively smaller and the chance of a rebounding relatively greater, the more violently the atom is vibrating—then by this theory the mean free path would diminish as the temperature rises, which is what is desired. This is an idea proposed long since by Wien.

The new idea is in result the same. The probability of the rebounding, or let me say of the *scattering* of an electron by an atom, is supposed to increase with the vigor of the vibrations of the latter. But for this a new reason is advanced: the reason, that while vibrating the atom is most of the time away from its equilibrium-place in the crystal lattice, and its relations with its companions are distorted. The probability of scattering is made to depend not only on the presence of the atom somewhere along the path down which the electron is rushing, but also on the relative positions of all the other atoms in the crystal.

To make such an assumption is, in effect, to compromise between the corpuscle-theory and the wave-theory. For what is the evidence from which

[13] The value $1/N\pi R^2$ familiar in the kinetic theory of gases, N standing for the number of fixed spheres per unit volume, R for the sum of the radii of a fixed and a moving sphere.

it is inferred that a beam of light falling upon a grating, say, or a beam of x-rays falling upon a crystal, are undulatory? Essentially this: the way in which the beam is scattered or diffracted by the regular array of rulings on the grating or atom-groups in the crystal is different—greatly and strikingly different—from the way in which we know that it would be scattered by a single ruling, or suspect with good reason that it would be scattered by a single atom-group. For instance, there are directions in which no light at all is sent by the regular array, though assuredly light would be scattered in those directions by any member of the array if it were solitary. These facts are explained by invoking interference of waves. The wavelets expanding outwards from the various rulings or scattering particles are supposed to arrive in opposite phases at the "dark fringes" of the diffraction-pattern, so that they cancel each other. But one might also say that the beam of light is a stream of corpuscles which are deflected or scattered by the atom-groups or rulings which they happen to strike, and that the law of scattering of the individual atom-group is altered by the marshalling of the scattering elements into a regular pattern, so that in particular the probability of a deflection towards one of the dark fringes is reduced to zero.

I am not prepared to say that such a compromise is a full alternative for the wave-theory, though modern theoretical physics seems to be tending in that direction. But if we wish to describe with the language of the corpuscle-theory the phenomena of diffraction by a crystal, whether of waves of light or waves of negative electricity: then we must certainly adopt the idea of a probability-of-scattering, of a mean-free-path, which varies with the irregularity of the placing of the atoms.

The principle is especially simple and especially startling, if we deal with a beam of which the wave-length—considering it as a beam of waves—exceeds the spacing of the lattice. Waves of such a magnitude would not be diffracted at all by scattering particles placed exactly at the points of the lattice. Though any particle singly would scatter them, they flow through the lattice intact. If then we wish to interpret the beam as a stream of corpuscles, the probability of deflection of any corpuscle by any atom must sink to zero when the arrangement is made perfect; the mean free path must then be considered infinite.

The resistance of a perfect crystal of an element should then be zero when all the atoms are stationary in their places on the lattice—if they ever are, which apparently they are not; and should increase steadily with increasing temperature, in a way which can be computed if we know two things: the way in which the scattering of waves by particles on a lattice varies with the amplitude of the quiverings of the particles about their lattice-points, and the way in which the amplitudes of the particles vary with the temperature. The second of these questions is the subject of the theory of specific heats of solids, developed principally by Debye. The first has been profoundly studied by Debye and by several other physicists interested chiefly in the scattering of x-rays by crystals. Transferring their

results into the theory of the diffraction of electron-waves, Houston demonstrated that over a wide range of temperatures the resistance of a perfect crystal should vary as the absolute temperature.

To determine not only the law of variation of resistance with temperature but the actual value of the resistivity for any metal, it would be necessary to evaluate the mean free path of the electrons. By the thoroughgoing corpuscular theory, this depends on the size of the atoms from which the corpuscles rebound; by the wave-theory, it depends on the scattering-power of the individual atom, which thus takes the place of the "size of the atom." The problem of computing the scattering-power of an atom for electron-waves belongs to the new mechanics. Houston was able to obtain good numerical agreements for several metals.

Another way of introducing irregularity into a crystal of an element consists in replacing a small fraction of the atoms, chosen at random here and there on the lattice, by atoms of another element. Certain alloys, known as "solid solutions," are of this type; and it is not only known that the resistance of such an alloy is greater than that of the element which is most abundant in it, but it has been shown by Nordheim that the dependence of resistance on percentage of substituted atoms follows the rule to be expected from the diffraction theory of resistance.[14]

Since then the conception of mean-free-path can be reinterpreted in terms of the wave-theory, and since it appears to be possible to deduce from the wave-theory a law of variation of mean-free-path with temperature which can be incorporated intact into the corpuscle-theory it is permissible to return to the corpuscle-picture to set up a theory of conduction of heat and of electricity, and of the thermo-electric effects in crystals.

We shall apply what I may call the method of the *perturbed distribution-function*, developed by Lorentz. The idea is, to begin by deriving a distribution suiting the actual case. The functions which we have hitherto employed, that of Maxwell and that of Fermi, are isotropic; it is only in the combination $(\xi^2+\eta^2+\zeta^2)$, hereafter to be called v^2, that the velocity-components ξ, η, ζ appear in them.[15] These "standard" functions may be appropriate to a uniform metal in which the temperature and the potential are uniform. Evidently they are not appropriate to a metal in which there is an electric field, or a temperature-gradient, or which varies in its chemical nature from place to place, as might an alloy. If in such a case we orient the axis of x parallel to the gradient of the variable quantity—be it electric potential, temperature, or whatever else—we must expect ξ to enter differently into the distribution-function from η and ξ.

Various arguments show that as a rule the departures from the standard function must be rather small. Lorentz therefore postulated that in the presence of a gradient directed parallel to the x-axis, the actual distribution

[14] The idea that the free paths of electrons extend from one irregularity of the crystal to another was propounded before the advent of the wave-theory of negative electricity.

[15] I shall use the velocity-components hereafter in lieu of the momentum components, to conform with the custom

should differ from the standard function $f_0(v)$—this he of course assumed to be Maxwell's—by virtue of a small additive term, a new function of v multiplied by the velocity-component ξ:

$$f=f_0(v)+\xi g(v) \tag{92}$$

and he proceeded to determine the new function g by the condition that f should remain constant in time despite the collisions of the electrons with the atoms. More precisely, he found for each of the three cases with which we shall be concerned a function g, such that the distribution-function obtained by adding ξg to f_0 conforms to that condition. This justifies the procedure.

Much the simplest case of the three is that of a uniform metal at a uniform temperature, subject to an electric field; for here the distribution-function need not vary from place to place. It will be well to go through the reasoning in this instance, though the formula for conductivity in which it leads differs but little from (91).

It is required, to find a function g of the combination $(\xi^2+\eta^2+\zeta^2)^{1/2}$ such that if at any moment the distribution $(f_0+\xi g)$ prevails, it continues unchanged throughout time— the number of particles in any compartment or cell of the velocity-space (the momentum-space of the earlier pages, with its unit of length altered in the ratio $m:1$) stays constant. Choose a compartment enclosed between planes ξ and $\xi+d\xi$, η and $\eta+d\eta$, ζ and $\zeta+d\zeta$. Call it the compartment C. Its volume is $d\xi d\eta d\zeta$, which to save a profusion of Greek letters I will usually denote by $d\tau$. The number of particles in it is $f.d\tau$. This number must remain unchanged, though individual particles are constantly moving into and out of C in either of two ways—by "drift" and by "collision."

Owing to the field E, all of the particles have a steady acceleration $\alpha=eE/m$, because of which they are continually and continuously drifting from cell to cell. One easily sees that the number which thus drift out of C per unit time (it is best to think of "unit time" not as one second, but as a period small compared with the mean time between impacts) is equal to first approximation to $\alpha f(\xi+d\xi,\ \eta,\ \zeta)d\eta d\zeta$.[16] This loss is partly balanced by an inward drift of particles which are accelerated into the compartment from the one lying beyond the plane ξ; the balance is not perfect, for the number drifting in per unit time is equal to $\alpha f(\xi,\ \eta,\ \zeta)d\eta d\xi$ and there is a difference $\alpha(df/d\xi)d\tau$ outstanding.[17] This difference must be balanced by the entrances and exits of particles which undergo collisions.

Let adt represent the number of electrons which, being initially in the compartment C, suffer impacts during unit time and are thus suddenly

[16] During a time dt so short that αdt is small by comparison with $d\xi$, the particles which initially lie between the plane $(\xi+d\xi-\alpha dt)$ and the side $(\xi+d\xi)$ of the compartment move out of it, while the particles which initially lie between the side ξ of the compartment and the plane $(\xi-\alpha dt)$ move into it; these two classes of particles number $f(\xi+d\xi,\ \eta,\ \zeta)\ \alpha dt\cdot d\eta d\zeta$ and $f(\xi,\ \eta,\ \zeta)\ \alpha dt\cdot d\eta d\zeta$ respectively.

[17] This expression figures in the equations as a net loss, but in fact has a negative sign (since $df/d\xi<0$) and therefore is actually a gain.

bumped out of it; and bdt the number which, being initially in other compartments, suffer such impacts during unit time that they suddenly turn up in C. The function g must be so chosen, that the lack of balance between the electrons drifting out and the electrons drifting in is just compensated by the lack of balance between those bumped into the compartment and those which are bumped out:

$$\frac{eE}{m}\frac{df}{d\xi}=b-a. \tag{93}$$

We must therefore evaluate $(b-a)$ in terms of the distribution-function.

We already have a formula ready-made for the number of impacts experienced per unit time by the particles of speed v; it is v/l for each particle, so that:

$$a=(v/l)fd\xi d\eta d\zeta. \tag{94}$$

Since however we have also to compute b, we shall find it expedient to classify these impacts according to the destinations of the particles, so to speak—according to the compartments of velocity-space into which they are bounced. A particle of speed v is located, in the velocity-space, on a sphere of radius v centered at the origin. Collision with an immovable atom changes the direction, but not the magnitude of the velocity; in the velocity-space, the particle suddenly moves to some other point on the same sphere. When electrons jump out of the compartment C because of impacts, they land in the other compartments which with C form a spherical layer around the origin. When electrons jump into C because of impacts, they come from the other compartments of that same layer. We shall derive an expression for the number of particles leaping from C into any other cell C' of the layer, and an expression for the number leaping reversely. The difference or lack of balance between these numbers, integrated over all the cells C', will be the required quantity $(b-a)$.

We begin by inquiring how many particles make such impacts that their paths (in the coordinate-space, of course—not the velocity-space) are deflected through angles between say θ and $\theta+d\theta$. To be deflected through an angle θ, an electron must strike an atom at a point where its surface is so oriented, that the normal (which is the line of centres at the instant of collision) is inclined to the line of approach of the electron at the angle $\psi=\frac{1}{2}(\pi-\theta)$. Denote by R the radius of the atom, and suppose that the radius of the electron is negligibly small.[18] Think of all the $f\cdot d\xi d\eta d\zeta$ electrons which at some particular moment of time are in unit volume of the metal, and belong to the compartment C of the velocity-space. Imagine each of these to be the centre, in the coordinate-space, of a pair of circles lying in the plane perpendicular to its path, and having radii $R\sin\psi$ and $R(\sin\psi+$

[18] The formulae remain valid even if the diameter of the electron is supposed not negligibly small, provided that we interpret R as the sum of the radii of atom and electron; but the generalization is not, so far as I know, of any practical value.

$d\sin\psi) = R(\sin\psi + \cos\psi d\psi)$. As time goes on, let these circles travel in the direction normal to their plane with the speed v. During unit time each pair of circles traces out a pair of cylinders of length v, containing between them a cylindrical sheath of volume $v \cdot 2\pi R^2 \sin\psi\cos\psi d\psi$. Multiplying this by the number $f \cdot d\xi d\eta d\zeta$ of the electrons, we get the total volume included in all of these sheaths. Multiplying this by the number N of atoms in unit volume, we get the number of atoms located with their centres in these sheaths—which is the number of atoms so placed that in unit time, electrons of the stated cell impinge on them at angles between ψ and $\psi + d\psi$—which is the number of impacts per unit time in which electrons are deflected through angles between θ and $\theta + d\theta$, which accordingly is this:

$$N \cdot f d\xi d\eta d\zeta \cdot 2\pi v R^2 \sin\psi \cos\psi d\psi = f d\xi d\eta d\zeta \cdot N\pi R^2 v \cdot \tfrac{1}{2} \sin\theta d\theta \qquad (95)$$

It will be convenient to express this as a fraction of the total number of impacts—call it $Z d\tau$—experienced per unit time by all the electrons in question, which by integrating (95) is found to be:

$$Z d\tau = \pi N R^2 v \cdot f \cdot d\tau \qquad (96)$$

substituting which into (95) we get:

$$Z \cdot 2 \sin\psi \cos\psi d\psi = Z \cdot \tfrac{1}{2} \sin\theta \, d\theta. \qquad (97)$$

It will be observed that deflections smaller than 90° are equally numerous with deflections greater than 90°, so that on the average the electrons after impact have no reminiscence of their prior direction of motion, as I mentioned earlier. Also, comparing (96) with (94), one derives the expression for mean free path,

$$l = 1/N\pi R^2 \qquad (98)$$

cited already in a footnote. To appreciate the most important feature of the expression (97) we must however return to the velocity-space.

In the velocity-space, the electrons of which the paths in coordinate-space are deflected through angles between 2θ and $2\theta + d2\theta$ execute leaps from the compartment C into other compartments of the spherical layer aforesaid, located on a certain region thereof. These occupy a belt or collar on the sphere, intercepted between two cones drawn with their common apex at the centre, their common axis pointing towards C and their apical semi-angles equal to θ and $\theta + d\theta$ respectively. Now the area of this belt is itself proportional to $\sin\theta d\theta$. This is very important: for it means that the electrons which are bounced out of C by collisions are sprinkled uniformly over all the rest of the sphere. More yet: it means that the electrons which are bounced out of *any* cell of the spherical layer are sprinkled uniformly through all the rest of the layer.

Consider then the interchange of electrons between two cells of the layer, say C at (ξ, η, ζ) with volume $d\tau$, and C' at (ξ', η', ζ') with volume $d\tau'$. The number leaping from C to C' is equal to the total number of impacts occur-

ring in C multiplied by the ratio which the volume of C' bears to the volume V of the layer. The number leaping from C' to C is equal to the total number of impacts occurring in C', multiplied by the ratio which the volume of C bears to the volume of the layer. The excess of the latter over the former is then:

$$Z(\xi',\eta',\zeta')d\tau'(d\tau/V)-Z(\xi,\eta,\zeta)d\tau(d\tau'/V) \tag{99}$$

which with the aid of (96) and (98) may be written thus:

$$\frac{vd\tau}{l}[f(\xi',\eta',\zeta')-f(\xi,\eta,\zeta)]d\tau'. \tag{100}$$

This is the quantity of which the integral with respect to ξ', η', ζ', extended over the spherical layer, is equal to $(b-a)d\tau$—the net rate at which compartment C gains particles through impacts.

Making Lorentz' postulate about the form of the distribution-function f, and remembering that throughout the spherical layer the combination $(\xi'^2+\eta'^2+\zeta'^2)^{1/2}$ is confined within a narrow range of values around v we obtain:

$$b-a=(v/l)\cdot g(v)\cdot \int(\xi'-\xi)d\tau'/V. \tag{101}$$

To effect the integration it is expedient to change over to polar coordinates in the velocity-space. Leaving the origin where it was, and directing the polar axis towards C, we make the radial coordinates and the colatitude-angle identical with our v and θ, and for the meridian-angle we use the symbol ϕ. Dividing up the layer into compartments by latitude circles and meridians, we have for any one of them:

$$d\tau'/V=(1/4\pi)\sin\theta d\theta d\phi \tag{102}$$

Consequently we obtain:[19]

$$b-a=(v/4\pi l)g(v)\cdot\int_0^{\pi}d\theta\sin\theta\int_0^{2\pi}d\phi(\xi'-\xi) \tag{103}$$

Everything therefore hinges on the evaluation of $(\xi'-\xi)$—the change in the x-component of velocity which the electron incurs when it leaps from C' to C—as a function of θ and ϕ. Now it may be shown without much difficulty,[20] that:

[19] Lest someone be disconcerted by the apparent difference between this equation and that given by Lorentz, I remark that I am using θ to designate an angle twice as great as the one which he denoted by θ.

[20] Let $\mathbf{v},\mathbf{v}'$ represent the vector velocities of the electron before and after impact, $\mathbf{c}_1$ the unit vector along the line of centres at the moment of impact. The components of $\mathbf{v}$ and $\mathbf{v}'$ along the line of centres are equal in magnitude and opposite in direction; the components perpendicular to the line of centres are equal in magnitude and direction. Writing these statements down in vector notation:

$$\mathbf{v}\cdot\mathbf{c}_1=\mathbf{v}'\cdot\mathbf{c}_1;\ \mathbf{v}-(\mathbf{v}\cdot\mathbf{c}_1)\mathbf{c}_1=\mathbf{v}'-(\mathbf{v}'\cdot\mathbf{c}_1)\mathbf{c}_1=\mathbf{v}'+(\mathbf{v}\cdot\mathbf{c}_1)\mathbf{c}_1,\text{ hence}$$
$$\mathbf{v}'-\mathbf{v}=-2(\mathbf{v}\cdot\mathbf{c}_1)\mathbf{c}_1=-2v\cos\psi\cdot\mathbf{c}_1;\ \xi'-\xi=-2v\cos\psi\cos\omega$$

$$\xi' - \xi = -2v \cos\psi \cos\omega = -2v \sin\tfrac{1}{2}\theta \cos\omega \qquad (104)$$

wherein ψ stands as before for the angle between the line of approach of the electron and the line of centres at the instant of impact, and ω stands for the angle between the line of centres and the axis of x. There is also a standard formula[21] relating $\cos\omega$ to ψ, ϕ and the angle $\beta = \arccos(\xi/v)$ between the initial path of the electron and the axis of x, as follows:

$$\cos\omega = \cos\beta \cos\psi + \sin\beta \sin\psi \cos\phi. \qquad (105)$$

We now have everything necessary to do the integration in equation (103), and we find:

$$b - a = (v\xi/l)g(v). \qquad (106)$$

Substituting this into equation (93), the condition that the distribution-function $(f_0 + \xi g)$ shall be stable by virtue of perfect balance between the rates at which electrons are shifted from compartment to compartment by the impacts and by the accelerating field, we get:

$$(eE/m)\frac{d}{d\xi}(f_0 + \xi g) = (v\xi/l)g. \qquad (107)$$

If the term $\xi g(v)$ is truly small by comparison with the term $f_0(v)$, we may neglect the second term on the left; and since $(df_0/dv) = (df_0/dv)\ (dv/d\xi) = (\xi/v)\ (df_0/dv)$, the culmination of all the argument is in the formula:

$$\xi g(v) = \frac{\xi l}{v^2}\ \frac{eE}{m}\ \frac{df_0}{dv} \qquad (108)$$

for the alteration which the applied electric field imposes on the distribution. Notice that g involves ξ and η and ζ only in the combination v; this justifies the procedure of Lorentz.

Now each electron which during unit time crosses any surface imagined in the metal contributes an amount e to the current through that surface; but the contributions made by electrons crossing in opposite senses are opposite in sign —what we perceive as current is *net* current, the excess of the flow of charge one way over the flow the other. Conceive a plane surface-element of area da, normal to the field, therefore normal to the axis of x. We must classify the electrons which traverse it according to their values of ξ. Let $H(\xi)d\xi da$ represent the number which pass through in unit time, having at the moment of passage x-components of velocity in the range

[21] Imagine two planes P_1 and P_2 intersecting along a vertical axis, the dihedral angle between them being ϕ. Through a point O on the axis draw a horizontal plane N, and from O draw two lines of unit length OR_1 and OR_2, the former in plane P_1 and inclined at β to the vertical axis, the latter in plane P_2 and inclined at ψ to the vertical axis; ω is the angle between them. The points R_1 and R_2 are at heights $\cos\beta$ and $\cos\psi$ above the plane H. On the vertical line through R_2 put a point R_3 at distance $\cos\beta$ above H. Expressions for the sides of the right-angled triangle $R_1R_2R_3$ are easily obtained, and (214) is derived by applying the theorem of Pythagoras to them.

$d\xi$ at ξ. This is equal to the number which at any instant have their x-components of velocity in this range, and are situated in the right prism having da for its base and extending a distance ξ down the direction of x:[22]

$$H(\xi)d\xi da=\xi da d\xi \int_{-\infty}^{\infty} d\eta \int_{-\infty}^{\infty} d\zeta f(\xi,\eta,\zeta) \tag{109}$$

Since for electrons crossing in opposite senses ξ is of opposite signs, the integral of this expression over all values of ξ, multiplied by e, gives the net current through da when the proper distribution-function is inserted. Evidently the integral will vanish when f is isotropic; there are enormous flows of charge both ways through da, but they are balanced. Unbalance is brought about by the non-isotropic perturbation-term in the distribution-function. Making the postulate of Lorentz, we obtain for the net current through unit area perpendicular to the field, the current-density J, the expression:

$$J=e\int H(\xi)d\xi=e\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\xi d\eta d\zeta\cdot\xi^2 g(v). \tag{110}$$

Set for $g(v)$ the expression in (217), and to effect the integration transform into polar coordinates in the velocity-space, orienting the polar axis along the axis of ξ; integrating over the angles, one obtains:

$$J=\frac{le^2E}{m}\,\frac{4\pi}{3}\int_0^{\infty} v^2\frac{df_0}{dv}dv. \tag{111}$$

The final integration is easy if one chooses for f_0 the Maxwell function, not quite so easy if one chooses that of Fermi. A further step will be of some advantage. Integrating by parts, and noticing that v vanishes at the lower and f_0 at the upper limit, we find:

$$J=-(8\pi le^2E/3m)\int_0^{\infty} f_0 v dv \tag{112}$$

and the integral remaining, divided by n the number of electrons per unit volume, is seen to be $1/4\pi$ times the mean value of v^{-1},—the average of the reciprocal of the speed of the electrons, in the absence of the field. Denoting this by $\overline{v^{-1}}$, we may write as the general formula for conductivity:

$$\sigma=\frac{J}{E}=\frac{2}{3}\,\frac{le^2n}{m}\overline{v^{-1}}. \tag{113}$$

The analogy with (91) is obvious, but we must not be misled into identifying the average of the reciprocal speed with the reciprocal of the average speed; they are not quite equal.

[22] This would be immediately obvious, if all the electrons were moving parallel to the x-axis and made no impacts. Electrons having y and z components of velocity in addition to the x-component will drift obliquely out of the prism, and electrons making impacts will be thrown out of the range $d\xi$; but each electron thus lost will be balanced by another coming in from outside.

The actual final formulae obtained by the old and the new statistics —substituting, that is to say, the appropriate values of $\overline{v^{-1}}$ from (72b) and (72a, and putting $G=2$—are as follows:

$$\sigma=\frac{4}{3}\frac{e^2ln}{(2\pi mkT)^{1/2}} \quad \text{(old)} \tag{114a}$$

$$\sigma=\frac{8\pi}{3}\frac{e^2l}{h}\left(\frac{3n}{8\pi}\right)^{2/3} \quad \text{(new)} \tag{114b}$$

As Sommerfeld has shown, all of the reasoning by which (108) was reached remains intact even when it is supposed that l is a function of v; in that case, l remains under the integral sign in (111), and the integral itself is equal to $n/4\pi$ times the mean value of $v^{-2}d(lv^2)/dv$. This generalization may be of some value.

Uniform Metal with a Temperature-Gradient; Thermal Conduction.

We have now to find a function g such that the distribution $(f_0+\xi g$ is stable in a metal in which there is a constant gradient of temperature along the axis of x. When we find it, we shall be able to evaluate the integral

$$W=\tfrac{1}{2}m\int\int\int d\xi d\eta d\zeta\cdot g(v)v^2\xi^2 \tag{115}$$

which is like the integral in (219) except for the differently-chosen form of g and the substitution of $\frac{1}{2}mv^2$ for e, and therefore represents the net rate of flow of kinetic energy borne by electrons through unit area perpendicular to the gradient—the contribution of the electrons to the flow of heat, under the circumstances stated.

The standard distribution-function f_0, involving as it does the temperature T, is in this case itself a function of x. One might expect that this dependence of f_0 on x would be sufficient to fulfil the requirements. An isotropic distribution which varies from point to point cannot however be stable; the particles conforming to such a one at any given moment would proceed to drift off down the gradient. A stable distribution cannot be isotropic. We must repeat the process of balancing the rates at which particles enter and leave each compartment of the phase-space through collisions and through drift. I say the phase-space now, instead of the velocity-space; for this case is made more complicated than the previous one by reason of the fact, that we now must make the balance separately for the electrons contained in each of the six-dimensional cells $d\xi d\eta d\zeta dxdydz$, whereas previously we could make it *en bloc* for all of the particles in the entire metal comprised within any velocity-cell $d\xi d\eta d\zeta$.

Consider then the six-dimensional cell $d\xi d\eta d\zeta dxdydz=ds$, and the $f(\xi, \eta, \zeta, x, y, z)ds$ electrons in it. The first three factors in ds denote the range of velocity, the last three the range in position, within which an electron must lie if it is to belong to ds. Electrons in the proper range of position are con-

tinually entering or leaving the proper range of velocity, because of impacts. The net rate at which ds gains electrons in this way is given, as before, by (106). Electrons in the proper range of velocity are continually drifting into the proper range of position, coming into $dxdydz$ from the region adjacent to it on the side towards smaller or greater values of x, according as ξ is positive or negative. By the same sort of reasoning as led to the term $\alpha df/d\xi$ in (93), one sees that the net rate at which ds loses electrons in this way is $\xi df/dx$. Equating the two, we have:

$$\xi df/dx = b - a = (\xi v/l)g(v) \tag{116}$$

and when we put f_0 for f as before, on the ground that the term (ξg) in the full expression for f makes but a small contribution to the left-hand member, we have all that is required for computing g from (116) and W from (115) for whatever standard function we elect.

At this point, however, there arises a difficulty. If having adopted this way of determining g we proceed to compute the electric current J in the metal by formula (110) we find that it is not zero. The reasoning has led to the conclusion that wherever there is a net current of heat in a metal, there is also a net current of electricity. This conclusion is not in accord with experiment. Yet there is apparently no other way to circumvent it, than to suppose that when a gradient of temperature is maintained in a metal there arises a spontaneous internal electric field, of just such a magnitude as to counteract the electric current which would otherwise persist. The gradient of temperature calls forth a gradient of potential; the actual distribution-function is the one which is stable under both these gradients combined. In the bookkeeping of the compartment ds, the net gain from impacts $(b-a)$ is balanced against the sum of the net loss through drift in the coordinate-space $(\xi df/dx)$ and the net loss through drift in the velocity-space $(\alpha df/d\xi)$. Putting these statements into the form of equations, and denoting by E the hypothetical electric field and by $\alpha(=eE/m)$ the acceleration which it imparts to each electron, we have:

$$\alpha(df/d\xi) + \xi(df/dx) = b - a = (\xi v/l)g \tag{117}$$

$$J/e = \int\int\int d\xi d\eta d\xi \cdot \xi^2 g = 0 \tag{118}$$

a pair of equations for determining E and the function g.

Lorentz, adopting the Maxwell-Boltzmann function for f_0, solved the equations, and obtained:

$$W = \frac{8}{3} \frac{nlk^2T}{(2\pi mkT)^{1/2}} \left(\frac{dT}{dx}\right) \qquad \alpha = \frac{1}{2} \frac{k}{m} \frac{dT}{dx}. \tag{119}$$

Sommerfeld adopted the Fermi function, and obtained for the degenerate case:[23]

[23] To derive this formula it was necessary to proceed to the second-approximation expression for the Fermi distribution-function (equation 107a); the first approximation (equation 107b) merely yielded zero for W.

$$W=\frac{8\pi^3}{9}\frac{lk^2T}{h}\left(\frac{3n}{8\pi}\right)^{2/3}\left(\frac{dT}{dx}\right). \tag{120}$$

The coefficient of dT/dx in these expressions for W is by definition the *thermal conductivity*, usually denoted by κ. One notices that these expressions for κ like those for σ, involve the more or less disposable constants n and l. This however is not true of the ratio of the conductivities.

The Wiedemann-Franz Ratio.

For the ratio of thermal to electric conductivity, the old statistics and the new supply expressions involving nothing but T and the ratio of the universal constants k and e, and differing only by a slight numerical factor:

$$\kappa/\sigma=2(k/e)^2T(=4.2\times10^{-11} \text{ at } T=291°K) \tag{121}$$

by the old statistics, and

$$\kappa/\sigma=\tfrac{1}{3}\pi^2(k/e)^2T(=7.1\times10^{-11} \text{ at } T=291°K) \tag{122}$$

by the new.

This "Wiedemann-Franz ratio" seems to have been predestined to encourage the devotees of the electron-gas theory. Every other formula offered by the theory contained either n or l or both, and therefore could not serve as an ultimate critical test; for any discrepancy with the data could be removed by adjusting these constants. True, the ensemble of the formulae provided by the classical theory ran counter to the data in so many different ways, that the net result was quite unfavourable; but one could not point out any single prediction which was certainly wrong. If however the Wiedemann Franz ratio had departed by an order of magnitude or more from the value of $2(k/e)^2T$, the electron-gas theory could hardly have survived the blow. But in this one case where disagreement would have been fatal, there was agreement; not perfect, but rather too good to be discarded as fortuitous. For many of the familiar metals the ratio, when measured at room-temperature, turned out to be around 6 or $7\cdot10^{-11}$. This more than any other one fact was what kept alive the feeling, that in spite of all its difficulties the electron-gas theory must be fundamentally right.

For the twelve metals Al, Cu, Ag, Au, Ni, Zn, Cd, Pb, Sn, Pt, Pd and Fe, the average of the values of κ/σ at 291°K is $7.11\cdot10^{-11}$. The agreement with the prediction of the new statistics is more than good. It is so very good, that it must be partly accidental, especially as the individual values from which the average is formed depart from it by varying amounts. One may still doubt whether it is to be admitted as one of the items which compel the adoption of the new statistics. Drude, be it recalled, obtained the value $6.3\cdot10^{-11}$ out of the crude assumption that all of the electrons in any volume-element have the same speed.[24] It used to be regarded as rather amusing

[24] Drude of course could evaluate the ratio k/e without knowing either k or e accurately or at all, since it is the same as the ratio $N_0k/N_0e=N_0$ standing for the number of molecules in a gramme-molecule, the Loschmidt number—and N_0k is the gas-constant R while N_0e is the Faraday constant of electrolysis.

that the elaborate calculations of Lorentz merely impaired the agreement which Drude had attained in a naively simple way.

Both theories require that the ratio be proportional to T; this is fairly well satisfied over wide ranges of temperature, but at extreme degrees of cold there is marked divergence, which is inconvenient. It may be desirable to invoke other mechanisms of conduction to supplement the free electrons—as for instance the passing-along of electrons from atom directly to atom to assist in the conduction of electricity, or the transmission of elastic vibrations to aid in the transfer of heat. Indeed, when one reflects that insulators though they lack free electrons yet have some device for the transmission of heat, one wonders why this device should not be available to metals also, and exalt their values of κ and of κ/σ above the predictions of the electron-gas theory.

Intrinsic Potential Difference.

We have seen that in a metal where there is no electric current and yet there is a current of heat, an internal electric field must be imagined. We shall now see that in a metal where there is no electric current, but the number of electrons per unit volume varies from point to point, there must also be an internal electric field. This sounds plausible to intuition, for one would expect the electrons to diffuse from regions of higher to regions of lower density unless they were impeded by some force. The equations (117) and (118) enable us to evaluate this force.

Returning to these equations, introduce polar coordinates v, θ, ϕ in the velocity-space as we have formerly done; then $\xi = v\cos\theta$. Multiply both sides of equation (117) by $\cos\theta$; the right-hand member of the new equation is then proportional to $\xi^2 g$. Integrate both members of this new equation over the entire velocity-space. The integral on the right then vanishes by reason of (118), and for the integrals on the left we have:

$$\alpha \int d\tau (df_0/d\xi) \cos\theta + \int d\tau \xi (df_0/dx) \cos\theta = 0 \tag{123}$$

By obvious transformations we get:

$$\alpha \int d\tau \frac{df_0}{dv} \cos^2\theta + \frac{d}{dx} \int d\tau f_0 v \cos^2\theta = 0 \tag{124}$$

Integrating over the angles:

$$\frac{4\pi}{3}\alpha \int \frac{df_0}{dv} v^2 dv^3 + \frac{4\pi}{3} \frac{d}{dx} \int f_0 v^3 dv = 0. \tag{125}$$

Leaving the second term as it is, but integrating the first by parts, we find that as f_0 vanishes (whichever statistics we use) at one limit and v at the other limit of integration, we get:

$$-\frac{2}{3}\alpha\int 4\pi f_0 v^{-1}\cdot v^2 dv+\frac{1}{3}\frac{d}{dx}\int 4\pi f_0 v\cdot v^2 dv=0. \tag{126}$$

The integrals are written in this curious fashion, to bring out the feature that they are proportional to the mean values of functions—the functions v^{-1} and v, respectively—averaged over the electrons in question; which is to say, all the electrons contained in the compartment *dxdydz* of coordinate-space, to which equation (117) has reference. They are in fact equal to the products of these mean values, to wit the *mean reciprocal speed* and the *mean speed*, by the number *ndxdydz* of the electrons in the compartment *dxdydz*. Rewriting (125) accordingly, with overlinings to signify averages, and dividing out the factors *dxdydz* and 1/3, we get:

$$-2\alpha n\overline{v^{-1}}+d(n\bar{v})/dx=0 \tag{127}$$

and this is the equation for the acceleration α or the accelerating field $E=m\alpha/e$, required to counteract the electric current which otherwise would be produced in the presence of the gradient $d(n\bar{v})/dx$ of the quantity $n\bar{v}$. This is the gradient which evokes the hypothetical electric field; gradients of temperature or of concentration act indirectly, by making $n\bar{v}$ vary.

With the classical statistics the development is extremely simple, for $\bar{v}$ depends on temperature only, while n may be varied at will. Heretofore we have tacitly assumed that n remains the same while T and therefore $\bar{v}$ vary along the axis of x, so that:

$$d(n\bar{v})/dx=nd\bar{v}/dx=n(d\bar{v}/dT)(dT/dx) \tag{128}$$

and the reader can verify the expressions for α and W given in (119) by starting from this point. But now we will assume that T and $\bar{v}$ remain the same while n varies along the x-direction with a gradient dn/dx. Then:

$$d(n\bar{v})/dx=\bar{v}\cdot dn/dx. \tag{129}$$

Putting this into (127), and recalling that in the Maxwell distribution the mean values of v and v^{-1} are thus related,

$$\bar{v}=(2kT/m)\overline{v^{-1}} \tag{130}$$

one perceives that $\bar{v}$ disappears by division from the two sides of the equation, leaving this:

$$\alpha n=(eE/m)n=(kT/m)dn/dx \tag{131}$$

the desired equation for the necessary electric field. Integrating it, we obtain another of very familiar aspect:

$$n=n_0\exp(eE/kT)(x-x_0)=n_0\exp(-[V-V_0]/kT) \tag{132}$$

This is the celebrated equation of Boltzmann embodying the statement that if in an assemblage of particles at uniform temperature there are variations in the number-per-unit-volume from place to place, then there must also be a field of force against which work must be done to move a particle from place to place—and *vice versa*. Specifically: if at any two points P

and O the number-per-unit-volume of the particles has values n and n_0, there must be a field of force such that when a particle is moved from O to P its potential energy is increased by $-kT \cdot log(n/n_0)$. If the particles are electrons and the field of force is electric and derived from a potential having values V at P and V_0 at O, then of course this change in potential energy is expressed by $e(V-V_0)$.

Boltzmann's equation is so deeply rooted in modern physics, that it seems strange and suspicious that the new statistics should substitute another but it does. The reason for the innovation stands out very clearly in (127) when the absolute-zero extreme of the Fermi distribution is applied. Owing to the dependence of $\bar{v}$ on n, *owing to the interrelation between average speed and number per-unit-volume* which distinguishes a system conforming to the new statistics, the second term in (127) is no longer proportional to dn/dx. Instead, we have:

$$n\bar{v} = \frac{3}{4} n v_m = \frac{3hn}{4m}\left(\frac{3n}{4\pi G}\right)^{2/3} \quad ; \quad \overline{v^{-1}} = 3/2v_m \tag{133}$$

substituting which values into (237), we obtain:

$$eE/m = \alpha = \frac{h^2}{3m^2}\left(\frac{3}{4\pi G}\right)^{2/3} n^{-1/3} dn/dx \tag{134}$$

and integrating:

$$-E(x-x_0) = +(V-V_0) = \frac{h^2}{2me}\left[\left(\frac{3n}{4\pi G}\right)^{2/3} - \left(\frac{3n_0}{4\pi G}\right)^{2/3}\right]. \tag{135}$$

This is the formula which supplants Boltzmann's equation.

Consulting (71), we see that (135) may be rewritten thus:

$$e(V_1 - V_0) = (W_i)_1 - (W_i)_0 \tag{136}$$

which is to say: if there is equilibrium between two samples of electron-gas, both being at absolute zero and distributed according to the Fermi law, and the fastest electrons of the two having values of kinetic energy W_{i1} and W_{i0} respectively—then there is a potential-difference between the two, such that if the fastest electron of either group were to cross over to the other, its kinetic energy on arrival would be equal to that of the fastest electron of the group which it joins. So stated, the proposition is easy to remember, and one might even come to think it obvious.

Consider now a pair of pieces of different metals, in contact with one another. One may conceive that they are welded together by an alloy in which the proportion of either varies continuously from zero to one hundred percent, if one feels the need for a mathematical continuity. If the two pieces were separate, the number of electrons per unit volume would probably not be the same for the two; certainly it is not the same if the number of electrons is equal to the number of atoms per unit volume, for this varies from metal to metal. If the process of welding the metals together does not alter the concentration of the electrons in either at points remote

from the junction, then a potential-difference given by (132) or (135)—according as the old or the new statistics is the proper one—must arise between the metals. Taking Sommerfeld's example of potassium and silver: if in unit volume of each of these metals there are as many electrons as atoms, and if this state of affairs continued when the two are welded together, then between the interiors of the metals across the weld there must be a potential-difference of 4.2 volts,[25] potassium being negative. This figure is calculated by the formula (135); the classical formula gives a value considerably lower, about 0.04 volts. This contrast is characteristic. Both the new and the old statistics associate an internal or intrinsic potential-difference with a difference in electron-concentration, and *vice versa*; but the amount of the P. D. associated with a given pair of concentrations is by no means the same by the two theories; and in actual cases, the new statistics gives much the larger amount.

Though it is not actually possible to measure the potential in the interior of a metal, there are phenomena which indicate that between two metals

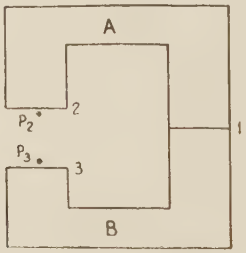


Fig. 2.

touching one another, or between two parts of a metal maintained at different temperatures, there is a difference of potential. These are the *thermoelectric* phenomena—Peltier effect, Thomson effect, thermal electromotive force. The internal potential-gradient reveals itself through the fact that when an electric current is sent through the region where it exists, the rate of generation of heat departs from that which is calculated by Joule's law. We must therefore apply the statistics—this will be the last application which I shall consider—to the problem of evaluating the transport and the generation of heat in an electron-gas, in which the distribution-function is perturbed by an electric field and simultaneously by either of the two other influences—varying temperature, varying concentration of electrons—which we have heretofore considered separately.

Before undertaking this I had better dispel any notion that the "contact" or "Volta" potential-difference between a pair of metals is the measure of the P. D. between their interiors for which we have just been deriving theoretical expressions. It is in fact a measure of something else, as one sees

[25] Sommerfeld originally computed 5.7 volts, having put $G=1$; the value 4.2 corresponds to $G=2$.

by examining an arrangement like that of Figure 2, where A and B signify pieces of two metals which are in contact at 1, and face one another across a gap between 2 and 3. Consider an electron anywhere inside A, and estimate the potential-barriers which it must cross in order to arrive at the point P_2 just outside of the boundary 2, and also those which it must cross in order to pass through the metal B and reach the point P_3 just outside of the boundary 3. Recalling the symbols and the relations introduced in the section on thermionics, one sees that there is a potential-difference between P_2 and P_3 given by the expression

$$(W_{aA}-W_{aB})-(W_{iA}-W_{iB})=b_A-b_B. \tag{137}$$

This is the contact potential difference; and we see that if the new statistics is correct, it is equal (at the absolute-zero limit) to the difference between the values, for the two metals in question, of that quantity b which appears in Richardson's equation and used to be regarded as the surface work-function. By the old statistics, it differs from (b_A-b_B) by the amount of the internal potential-difference between the metals across the junction 1. Perhaps this difference between the consequences of the two theories could be tested by experiment.

Theory of the Thermoelectric Phenomena.

We turn now to the problem of evaluating the rate of generation of heat in a metal through which an electric current is flowing, and in which (according to these theories) there is an intrinsic electric potential-gradient due to a temperature-gradient, or to a gradient of electron-concentration, or both together. The process leads to formulae which can be tested by experiment, furnishing thus some additional ways of finding out whether these ideas of the new statistics, of the perturbed distribution-function and of the internal electric field are justifiable.

The expression for the rate of generation of heat per unit volume in a conductor traversed by currents of electricity and heat flowing along the axis of x and having current-densities J and W respectively, is $(JE-dW/dx)$. Here E stands for the electric field— not in general for the applied electric field alone, but for the sum of this and the hypothetical internal field. I denote the corresponding acceleration by α, as before, and the rate at which heat is generated per unit volume by r; then:

$$-r=(m\alpha/e)J-dW/dx. \tag{138}$$

The current-density of heat is given by the formula (115), which I repeat:

$$W=\tfrac{1}{2}m\int v^4\cos^2\theta g d\tau. \tag{139}$$

Here g stands as always for the non-isotropic perturbation-term in the distribution-function. This and the acceleration α are to be determined from the two equations,

$$\alpha(df_0/d\xi)+\xi(df_0/dx)=(v\xi/l)g=(v^2\cos\theta/l)g \tag{140}$$

$$J/e=\int \xi^2 g d\tau=\int v^2\cos^2\theta g d\tau \tag{141}$$

which are the same as (117) and (118) except that the electric current is no longer set equal to zero.

Multiply both sides of (140) by $\frac{1}{2}mv^2\cos\theta$, and integrate over the entire velocity-space. The integral of the right-hand member is W/l; developing the integral of the left-hand member, we find:

$$W=\tfrac{1}{6}ml(-4\alpha n\bar{v}+d(n\overline{v^3})/dx) \tag{142}$$

Multiply both sides of (140) by $\cos\theta$, and integrate over the entire velocity-space; the integral of the right-hand member is J/le; developing the integral of the left-hand member, we get the equation of which (127) was a special case, to wit:

$$-2\alpha n\overline{v^{-1}}+d(n\bar{v})/dx=3J/el \tag{143}$$

Evidently these equations suffice to translate (138) into an expression for r in terms of the mean free path, the universal constants, and the averages of various powers of v.

Postulating the Maxwell-Boltzmann distribution-function for f_0; working out the expressions for α and for dW/dx, and importing the formulae for σ (equation 114a) and κ (equation 119), one finds:

$$J(m\alpha/e)=J\cdot\frac{1}{2}\frac{k}{e}\frac{dT}{dx}-\frac{J^2}{\sigma} \qquad dW/dx=\frac{d}{dx}\left(\kappa\frac{dT}{dx}\right)+2J\frac{k}{e}\frac{dT}{dx} \tag{144}$$

so the value predicted for the rate of generation of heat per unit voume amounts to this:

$$r=+\frac{J^2}{\sigma}+\frac{d}{dx}\left(\kappa\frac{dT}{dx}\right)+\frac{3}{2}\frac{k}{e}\frac{dT}{dx}J. \tag{145}$$

The first term is obviously the Joule heat; the second is not directly a consequence of this current-flow, as it would occur whatever the agency which set up the temperature-distribution in question. It is the third term which concerns us; this is a "reversible heat," proportional to the first power of the current, so that when the current flows in one sense heat is absorbed and when it is reversed heat is evolved. The sign is such, that heat is absorbed when the electrons are flowing towards the hotter part of the metal; the magnitude is such, that as the electrons move onward they acquire just enough energy to raise their temperature to that of the regions which they enter. The coefficient of this term therefore represents the specific heat of the electron-gas, which is the same as that of any other monatomic gas when referred to equal numbers of particles.

Adopting instead the Fermi distribution, and inserting into (142) and (143) the values of $\bar{v}$ and $\overline{v^{-1}}$ and $\overline{v^3}$ prevailing at absolute zero, we find on making the substitutions in the expression (138) for r that the terms con-

taining the first power of J balance one another out. This might have been expected; for we have just seen that these terms form a sum which is proportional to the specific heat of the electron-gas; and if this result may be extended to an electron-gas conforming to the Fermi distribution, then since the specific heat vanishes at zero so also must this "reversible heat." Working through the second approximation, Sommerfeld found that the net coefficient of the term in J in the expression for r is in fact proportional to the specific heat of the electron-gas, being therefore proportional to T, and given by the formula:

$$\frac{2\pi^2}{3}\frac{mk^2}{eh^2}\left(\frac{2\pi G}{3n}\right)^{2/3}T. \qquad (146)$$

Now it is a fact of experience that when an electric current flows along a uniform wire of uneven temperature, heat is generated at a rate which involves a term proportional to the current and which changes sign when the current changes sense. This "Thomson heat," like the maximum value which experiments allow us to admit for the specific heat of the electron-gas, has always been much smaller than the value which the classical statistics requires provided that the free electrons are about as numerous as the atoms. For the Thomson heat as for the specific heat, the new statistics sharply reduces the amounts demanded—to about one percent of those on which the classical theory insists, at room-temperature that is to say and assuming always that the free electrons are equal in number to the atoms. Agreement in order of magnitude is now attained, and for some metals the advantage is possibly greater; there are indications, too, that the Thomson heat is proportional to T over wide ranges of temperature.

Finally we consider the "Peltier heat"——a term proportional to the current and changing sign when the current changes sense, observed when there is a flow of electricity across a weld or area of contact between two metals. This is clearly to be interpreted as a term in the first power of J, occurring when into the combination $J(m\alpha/e)$—all that remains of the expression (138) for r, when the gradient of temperature is annulled —we substitute the value of α derived from (143) with the assumption that n varies continuously across the weld from the value appropriate to the one metal to the value appropriate to the other. Using the classical statistics, we find that there is such a term; denoting by n and n_0 the electron-concentrations in the two metals, we find for its value:

$$(kT/e)\log(n/n_0)J. \qquad (147)$$

Its value for unit current is obviously the intrinsic potential-difference be- the metals. Using instead the new statistics, we find that at the absolute zero there is no such term; we must proceed to the next approximation, doing which, Sommerfeld obtained the expression:

$$\frac{2\pi^2}{3}\frac{m(kT)^2}{eh^2}\left[\left(\frac{4\pi G}{3n}\right)^{2/3}-\left(\frac{4\pi G}{3n_0}\right)^{2/3}\right]. \qquad (148)$$

Putting the current equal to unity, we find a value very considerably smaller than the intrinsic potential-difference between the metals—a fraction of a millivolt. This is the order of magnitude of the Peltier heat as it is actually observed in many cases. Curiously enough, this fact by itself is in accord with both the theories. By the classical statistics, the intrinsic potential-difference between two metals is generally small, and the Peltier heat for unit current gives its value directly; by the Fermi statistics, the intrinsic potential-difference is generally large, but the Peltier heat for unit current is only a small fraction of it.

Omissions

Among the subjects omitted from this article there are several of much interest, which the reader may trace from the annexed bibliography; in particular:

Sommerfeld's theory of the Hall effect;

Houston's extension of Sommerfeld's theory of intrinsic potential difference, including especially an explanation of the Peltier heat arising when current flows between two differently-oriented crystals of a single substance;

Bloch's use of the new methods of quantum mechanics to make allowance for the influence of the atoms on the conduction-electrons;

Fermi's application of statistical methods to the problem of determining the distribution of electrons in the individual atom;

Fuerth's work on the fluctuations in the new statistics.

The bibliography will indicate other interesting advances in a variety of problems.

Bibliography

J. M. Adams: Value of W_0 for silver. Zeits. f. Physik **52,** p. 882 (1929).

W. Anderson: Zeits. f. Physik **50,** p. 874–877 (1928). Astrophysical applications.

H. M. Barlow: Phil. Mag. (7) **7,** pp. 458–470 (1929). A criticism of the electron theory of metals.

R. S. Bartlett: Proc. Roy. Soc. **A121,** pp. 456–464 (1928). Thermionic and cold discharge.

E. S. Bieler: Journ. Franklin Inst. **206,** pp. 65–82 (1928). General account of new statistics of electron-gas.

H. F. Biggs: Nature **121,** pp. 503–506 (1928). General.

F. Bloch: Zeits f. Physik **52,** pp. 555–559 (1928). Quantum mechanics of electrons in crystal lattices.

F. Bloch: Susceptibility and change of resistance in magnetic field. Zeits f. Physik **53,** pp. 216–227 (1929).

Bose: Zeits. f. Physik **26,** pp. 178–181 (1924). First paper on the Bose statistics.

L. Brillouin: C. R. **184,** pp. 589–591 (1927).

P. A. M. Dirac: Proc. Roy. Soc. **A112,** pp. 661–677 (1926).

J. W. M. DuMond: Proc. Nat. Acad. Sci. **14,** pp. 875–878 (1928); Phys. Rev. (2) **33,** pp. 643–658 (1929). Scattering of x-rays by the free electrons.

C. Eckart: Zeits. f. Physik **47,** pp. 38–42 (1928). Contact P. D.

A. Einstein: Berliner Sitzungsberichte, 1924, pp. 261–267; 1925, pp. 3–17. Application of Bose statistics with modification to material gases.

E. Fermi: Zeits. f. Physik **36,** pp. 902–912 (1926). Fermi statistics.

E. Fermi: Zeits. f. Physik **48,** pp. 73–79 (1928). Application of statistics to electrons within atoms.

E. Fermi: Zeits. f. Physik **49,** pp. 550–554 (1928). Statistical evaluation of the *s*-levels.

E. Fermi: *Quantentheorie und Chemie* (Hirzel, 1928), pp. 95-111. Application of statistics to problems of atomic structure.
V. Fock: Zeits. f. Physik **49,** pp. 339–357, (1928). Dirac's statistical equation.
R. H. Fowler: Proc. Roy. Soc. **A117,** pp. 549–552 (1928). Thermionics.
R. H. Fowler, L. Nordheim: Proc. Roy. Soc. **A119,** pp. 173–181 (1928).
R. H. Fowler: Proc. Roy. Soc. **A120,** pp. 229–232 (1928). Photoeffect.
R. H. Fowler: Proc. Roy. Soc. **A122,** pp. 36–49 (1929). Thermionic constant A.
R. H. Fowler: Monthly Notices Royal Astron. Soc. **87,** pp. 114–122 (1926). Astrophysical application.
J. Frenkel: Zeits. f. Physik **49,** pp. 31–45 (1928). Theory of magnetic and electric properties of metals at absolute zero.
J. Frenkel, N. Mirolubow: Zeits. f. Physik **49,** pp. 885–893 (1928). Theory of Metallic Conduction by Wave Mechanics.
J. Frenkel: Zeits. f. Physik **50,** pp. 234–248 (1928). Cohesion.
R. Fuerth: Zeits. f. Physik **48,** pp. 323–339 (1928), and **50,** pp. 310–318 (1929). Fluctuations in the New Statistics.
G. E. Gibson, W. Heitler: Zeits. f. Physik **49,** pp. 465–472 (1928). Chemical Constant in the New Statistics.
E. H. Hall: Proc. Nat. Acad. Sci. **14,** pp. 366–380, 802–811 (1928). Arguments against the electron-gas theory.
W. V. Houston: Zeits. f. Physik **47,** pp. 33–37 (1928). Cold Discharge.
W. V. Houston: Zeits. f. Physik **48,** pp. 449–468 (1928). Conduction and Thermoelectric effects in Metals.
W. V. Houston: Phys. Rev. (2) **33,** pp. 361–363 (1929). Cold Discharge.
O. Klein: Zeits. f. Physik **53,** pp. 157–165 (1929). Reflection of Electrons at a potential-discontinuity by Dirac's Dynamics.
E. Kretschmann: Zeits. f. Physik **48,** pp. 739–744 (1928). Continuation of Sommerfeld's work.
J. E. Lennard-Jones: Proc. Phys. Soc. Lond. **40,** pp. 320–537 (1928). General Account.
J. E. Lennard-Jones, H. J. Woods: Proc. Roy. Soc. **A120,** pp. 727–735 (1928). Electron-gas theory with modifications to allow for atoms in metals.
A. C. G. Mitchell: Zeits. f. Physik **50,** pp. 570–576 (1928). Entropy of the electron-gas.
L. Nordheim: Phys. Zeits. **30,** pp. 177–196 (1929). General report on new theory of thermionic emission, cold discharge and photoeffect. Earlier papers: Zeits. f. Physik **46,** pp. 833-855 (1928); Proc. Roy. Soc. **A121,** pp. 626–639 (1928); **A119,** pp. 173–181 (1928) (with R. H. Fowler).
L. Nordheim: Proc. Roy. Soc. **A119,** pp. 689–698 (1928). Alternative way of deriving the distribution-laws.
L. Nordheim: Naturwiss. **16,** pp. 1042–1043 (1928). Conductivity of alloys.
J. R. Oppenheimer: Proc. Nat. Acad. Sci. **14,** pp. 363–365 (1928). Cold discharge.
W. Pauli: Zeits. f. Physik **41,** 81–102 (1927). Paramagnetism.
W. Pauli: *Probleme der modernen Physik* (Hirzel, 1928) pp. 30–45. H-theorem.
F. Rasetti: Zeits. f. Physik **49,** 536–549 (1928). Statistical evaluation of the M-levels.
O. K. Rice: Phys. Rev. (2) **31,** pp. 1051–1059 (1928). Electrocapillarity.
O. W. Richardson: Proc. Roy. Soc. **A117,** pp. 719–730 (1928). Cold discharge.
L. Rosenfeld, E. E. Witmer: Zeits. f. Physik **47,** pp. 517–521 (1928). The radiation-gas.
L. Rosenfeld, E. E. Witmer: Zeits. f. Physik **49,** pp. 534–541 (1928). Refractive indices of metals for electron-waves, values of W_a.
R. Ruedy: Phys. Rev. (2) **32,** pp. 974–978 (1928). Dispersion of metals.
E. Schroedinger: Phys. Zeits. **27,** pp. 95–101 (1927).
A. Sommerfeld: Zeits. f. Physik **47,** pp. 1–32, 43–60 (1928). The fundamental article on the application of the Fermi statistics to electrons in metals. Preliminary note in Naturwiss. **15,** pp. 825–832 (1927).
A. Sommerfeld: Naturwiss. **16,** pp. 374–381 (1928). Extension of the article just listed.
A. T. Waterman: Phil. Mag. (7) **6,** pp. 965–970 (1928). Dependence of conductivity on pressure.
A. T. Waterman: Proc. Roy. Soc. **A121,** pp. 28–41 (1928). Cold discharge.
G. Wentzel: *Probleme der modernen Physik* (Hirzel, 1928), pp. 79–87. Photoeffect.

VOLUME 1 OCTOBER, 1929 NUMBER 2

THE PHYSICAL REVIEW SUPPLEMENT

CONDUCTED BY THE
AMERICAN PHYSICAL SOCIETY

STUDIES IN CONTEMPORARY PHYSICS

BOARD OF EDITORS

JOHN T. TATE, University of Minnesota

A. H. COMPTON K. T. COMPTON K. K. DARROW
E. C. KEMBLE C. E. MENDENHALL D. L. WEBSTER

CONTENTS

General Principles of Quantum Mechanics. Part I . . EDWIN C. KEMBLE 157
Recombination and Photo-Ionization . . . . . F. L. MOHLER 216
Spectroscopy of Soft X-Rays . . . . . THOMAS H. OSGOOD 228
Errata 241

PUBLISHED QUARTERLY
EDITORIAL OFFICES: MINNEAPOLIS

THE AMERICAN PHYSICAL SOCIETY

OFFICERS OF THE SOCIETY

President:	HENRY G. GALE	University of Chicago, Chicago, Ill.
Vice-President:	W. F. G. SWANN	Bartol Research Foundation, Philadelphia, Pa.
Secretary:	W. L. SEVERINGHAUS	Columbia University, New York City
Treasurer:	G. B. PEGRAM	Columbia University, New York City
Editor:	JOHN T. TATE	University of Minnesota, Minneapolis, Minn.
Local Secretary for the Pacific Coast:	L. B. LOEB	University of California, Berkeley, Calif.

THE PHYSICAL REVIEW SUPPLEMENT

STUDIES IN CONTEMPORARY PHYSICS

THE PHYSICAL REVIEW SUPPLEMENT is published quarterly under the auspices of the American Physical Society. Its purpose is to provide those interested in developments in physics with comprehensive, authoritative and timely discussions of the problems which are, at the present time, of especial interest and importance.

It is intended that in so far as possible these discussions shall be written in such form as to interest and inform those who are not specialists in the subjects under discussion.

It is hoped that the studies appearing in the SUPPLEMENT will give to the specialist the benefit of the view point, in a broad sense, of other specialists; to the student, a background of critical knowledge and an adequate bibliography of source material; to those whose time is absorbed in the teaching of physics, a stimulating account of the forward movement of their science.

THE PHYSICAL REVIEW SUPPLEMENT is published quarterly in January, April, July and October by the American Physical Society at 450 Ahnaip St., Menasha, Wis. The volumes begin with the July number. The first volume began in July, 1929.

SUBSCRIPTION PRICE

To members of the American Physical Society or of the American Chemical Society	**per year**	**$3.00**
To all others	**per year**	**4.00**
Single copies		**1.25**

All correspondence should be addressed to THE PHYSICAL REVIEW SUPPLEMENT, University of Minnesota, Minneapolis, Minnesota.

Application has been made for entry as second class matter at the post office at Menasha, Wisconsin.

Volume 1 October, 1929 Number 2

THE
PHYSICAL REVIEW
SUPPLEMENT

THE GENERAL PRINCIPLES OF QUANTUM MECHANICS. PART I

By Edwin C. Kemble
Harvard University

SECTION 1. ELEMENTARY PRINCIPLES OF WAVE MECHANICS

1. Historical Introduction

FROM the empirical standpoint the most striking features in the development of twentieth-century physics have been (a) the discovery and exploitation of the corpuscular characteristics of radiation; (b) the discovery and exploitation of the existence of discrete atomic and molecular energy levels, bringing into mechanics a new kind of atomicity superposed on the atomicity of electron and proton; (c) the discovery of the wave-like characteristics of matter as exhibited in the now famous experiments of Davisson and Germer. The quantum mechanics is the fruit of the efforts of theoretical physicists to provide a concise and unified description of these new experimental facts and of the great mass of empirical data on which the "classical" physics of the nineteenth century was built.

Roughly speaking we may divide the history of twentieth-century theoretical physics into two periods. In the first period, extending from the initial discovery of the Planck radiation formula until 1925, the leaders in theoretical physics were Planck, Einstein, Bohr, and Sommerfeld. Einstein led the way in the formulation of the corpuscular theory of radiation which proved so stimulating to experimental physicists. The many successes of this point of view culminating in its interpretation of the Compton effect served to place it on an equal footing with the classical wave theory. The older theory was not completely overthrown, however, as it gave the only natural interpretation of the facts regarding interference and diffraction. In practice physicists were led to the adoption of a dualistic mode of thought in which they treated light as waves in the discussion of one class of experiments and as corpuscles in the discussion of another.

In parallel with the development of the corpuscular theory of radiation went the formulation of the Bohr quantum theory of atomic structure with its marvelously useful interpretation of line and band spectra and of the collision experiments of Franck, Hertz, K. T. Compton, and others. This theory performed most admirably the function of correlating the

empirical facts in the field which it covered and through the correspondence principle laid the foundation for the matrix mechanics of Heisenberg, Born, and Jordan. It was at all times, however, confessedly incomplete and in many respects inaccurate.

The second period witnessed the simultaneous development of the matrix mechanics and of the wave mechanics of de Broglie and Schrödinger. These theories, now fused together into what we call quantum mechanics, correct the deficiencies of the Bohr theory as a tool for investigating the structure of matter, remove its restriction to periodic motions, relate the diffraction of electrons to the problem of atomic structure, and throw a great deal of new light on the dilemma regarding the nature of radiation.

In the form of quantum mechanics now most widely used, the dualistic nature of radiation is accepted as a brute fact to be described rather than explained or exorcised. A similar dualistic nature is ascribed to matter and thus a unification in the treatment of matter and radiation is attained. The discovery of the fundamental similarity between matter and radiation is one of the most striking features of present physical theory. Differences, of course, remain and we can by no stretch of the imagination identify these two fundamental modes of existence, but the analogy is far-reaching enough to permit the use of observations regarding the characteristics of radiation as guides in the formulation of a theory of matter. The de Broglie-Schrödinger theory was the result of a conscious attempt to follow such guides and this method of approach to the quantum mechanics seems to the present writer to be more natural than any other.

2. The Dualistic Theory of Radiation

The importance of optical analogy in the development of quantum mechanics lies primarily in the fact that the dualistic nature of light is much more obvious than the dualistic nature of matter. In the region of long waves the wave-like characteristics of radiation are strongly predominant, while in the x-ray region the corpuscular characteristics are predominant. As the transition from the one part of the spectrum to the other is continuous, the dualism is inescapable.

Let us then begin our survey of the quantum mechanics with a brief preliminary examination of the problem of radiation. What is the upshot of the battle between the wave theory of light and the corpuscular theory? The answer is, "A deadlock."[1] The wave theory gives a simple and accurate account of interference, diffraction, and dispersion besides making proper connection with quasi-static electromagnetic phenomena in the limiting region of very long waves. The corpuscular theory gives a simple and accurate account of the fundamental laws of the photo-electric effect and the Compton effect; it is a logical corollary of the fundamental law of spectro-

[1] Here we explicitly ignore the recent "double quantization" theory of Jordan and his collaborators which achieves a genuine fusion of the two theories by treating the electromagnetic vectors E and H as matrix quantities. This development should logically follow that of the quantum mechanics.

scopy $E'-E''=h\nu$; and it accounts for the abrupt changes in momentum experienced by emitting and absorbing atoms and molecules in a radiation field. Neither point of view gives a satisfactory description of the whole field of optics. In the case of a few phenomena, notably the Doppler effect, the predictions of the two theories are identical and agree with experiment. In other cases the two theories supplement each other. For example, in the case of the inverse photo-electric effect or the production of the continuous x-ray spectrum, the corpuscular theory is needed to account for the sharply defined limit to the spectrum, but must be supplemented by the wave theory in order to take into account the polarization phenomenon. Either point of view gives a qualitative explanation of the variation in hardness with direction of emission, which is, in fact, a kind of Doppler effect.

Evidently we need a fusion of these two theories. Perhaps the most obvious mode of procedure for uniting the two points of view is to assume that light consists of "wave packets" or wave disturbances occupying a small volume in space. Such a fusion of the two theories meets insuperable difficulty, however, in the fact that wave packets except under very special circumstances spread out indefinitely as they progress, instead of remaining small. Moreover, the facts regarding interference and diffraction require spreading wave fields rather than wave packets. To see this, we observe first that to account for the passage of light through small holes and for the basic phenomena which have given rise to the corpuscular conception, we must assume that the wave packets are very small. Interference would have to be accounted for as due to the overlapping of different packets. But the experiments on interference in weak light show clearly that interference fringes are perfectly formed even when the photons are so far apart that such overlapping is out of the question.

A more successful way of uniting the two theories is to adopt a frankly dualistic point of view by assuming that light consists of *both* spreading waves and corpuscles. The familiar fundamental formulas of Einstein for the energy and momentum of a light corpuscle, i.e.

$$E=h\nu \tag{1}$$

$$p=h\nu/c=h/\lambda \tag{2}$$

point directly to such a dualism, since they are really merely a basic lexicon for translating wave language into corpuscle language. If we postulate such a dualism, we can suppose that the electromagnetic waves of Maxwell's theory act as guides to the corpuscles and thus produce interference phenomena,[2] while the localization of energy and momentum in the photons accounts for the photo-electric effect, etc.

Of course, the existence of two apparently conflicting sets of characteristics for radiation has been a commonplace for many years and to many physicists the adoption of a dualistic point of view as the starting point for a fresh attack on the fundamental problems of physics will seem an evasion

[2] Cf. W. F. G. Swann, Science, **61**, p. 433 (1925).

of the fundamental question, "*Why* does light act in some respects like an assemblage of corpuscles and in other respects like a spreading wave phenomenon?" We assert, however, that in the last analysis the function of theoretical physics is to describe rather than to explain. Science seeks to interpret the infinitely complex world of direct experience as the outcome of fundamentally simple laws. The reduction of complexity to simplicity is the goal, and when it is attained, we prove that order underlies chaos and leave the question, "Why" still essentially untouched. Hence, discarding this question as ultimately unanswerable, we may address ourselves to the task of describing what we observe in the most compact manner possible. If the behavior of radiation can be at least approximately described by means of the dualistic point of view, its temporary adoption will be a step in advance. No claim of ultimate validity is made for the theory, however.

An apparently essential feature of any such fusion of the wave and corpuscle theories is its statistical character. As the corpuscles are discontinuously distributed in space while the waves are continuous, it is evident that if the waves are to determine the positions of the corpuscles, they can only do so statistically. In other words we must regard the intensity of light of frequency ν in any small volume G as a measure of the probable number of photons of energy $h\nu$ in G. This mode of correlating the positions of waves and corpuscles we will carry over directly from the radiation problem to the matter problem.

3. An Analogy Between Geometrical Optics and Classical Mechanics

The wave or "optical" theory of matter has its roots in the formulation of the principles of least time for geometrical optics and of least action for Newtonian mechanics by Fermat and Maupertuis, respectively. The principle of least time states that the path of a ray of light (wave-front normal) from a point A to a point B is always such as to make the integral

$$\int_A^B \frac{ds}{w}$$

an extremal (usually a minimum) with respect to all other conceivable paths for rays of the same color or frequency. w in this formula denotes the phase velocity of light and is a function of the frequency and of the space coordinates x, y, z. As the frequency ν is treated as a constant in varying the integral, and as the wave-length λ is equal to w/ν, the principle may equally well be stated in the form

$$\delta \int_A^B \frac{ds}{\lambda} = 0. \qquad (\nu \text{ unvaried}) \quad (3)$$

This means that the path length measured in wave-lengths is an extremal. As the spreading of light waves in space is fully determined by the "wave equation"

$$\nabla^2\psi = \frac{1}{w^2(x,y,z,\nu)} \frac{\partial^2\psi}{\partial t^2} \tag{4}$$

we may conclude that Fermat's principle is deducible under suitable conditions from this equation. The proof is very simple in the case of plane waves moving in homogeneous media bounded by plane surfaces of discontinuity. The generalization to non-homogeneous media is not given in the classical treatises on optics, but de Broglie[3] has shown how to carry it through. For this purpose it is necessary to assume that the wave whose rays are to be determined is progressive and only slightly curved.[4] Also the fractional change in the phase velocity w or in the wave-length λ in the distance λ must be small. These restrictions create no difficulty in the formulation of a wave theory of matter.

Let us now compare Fermat's principle in the form of Eq. (3) with the principle of least action. In case of a single particle of total energy E, kinetic energy T, and mass μ, moving through a force field with potential energy $V(x, y, z)$ the latter principle requires that the action integral

$$S = \int_A^B 2T dt = \int_A^B [2\mu(E-V)]^{1/2} ds \tag{5}$$

over the natural or mechanical path between two points A and B shall be an extremal as compared with its value for all adjacent paths using the same value of E. For such a particle the integrand $[2\mu(E-V)]^{1/2}$ is identical with the absolute value of the instantaneous momentum p which the particle would assume at x, y, z so that the principle may be stated in the form

$$\delta \int_A^B p(E,x,y,z) ds = 0. \qquad (E \text{ unvaried}) \tag{6}$$

The analogy between these two principles was seized upon by Sir William Hamilton in the early part of the nineteenth century and used as a guide in the development of dynamical and optical theory. Hamilton regarded it as an analogy only, however, and it remained for de Broglie and Schrödinger to show that in view of this analogy Newtonian mechanics may be regarded as a limiting case of a suitably defined wave problem.

A comparison of equations (3) and (6) shows that the paths of particles in the Newtonian dynamics may be identified with the "rays" in a wave problem in which the index of refraction and phase velocity are made suitable functions of x, y, z, and the energy E. To be exact, the condition imposed on the wave problem is that

$$\frac{C}{\lambda} = \frac{C\nu}{w} = p(E,x,y,z) = (2\mu[E-V(x,y,z)])^{1/2}, \tag{7}$$

where C is an arbitrary constant.

[3] L. de Broglie, Journ. de Physique **7**, p. 321 (1926).

[4] Radius of curvature large compared with the wave-length.

4. Wave Packets and Group Velocity

We have set out to develop a theory of mechanics parallel to our theory of light. Hence we must demand *ab initio* that the relation between the motion of photons in optics and the motion of the associated optical waves shall be duplicated in the relation between the motions of large scale bodies and the associated "mechanical waves." Now the nearest that we can come to observing the motion of an individual photon is to form a "wave packet" or localized disturbance of an approximately monochromatic nature. To this end we may shine plane parallel monochromatic light on a small aperture provided with a shutter. If this be opened for a brief interval of time, it will allow a short train of waves to pass through. Such a train is called a wave packet, since it is a localized bundle of energy which may be resolved into a superposition of plane waves spread over a narrow range of directions and frequencies. The path of such a train may be approximately mapped out and its velocity can be approximately measured. In theory, at any rate, there is nothing to prevent the observation of the orbit of such a train even in an inhomogeneous dispersive medium. Of course, such a train will ordinarily contain more than one photon, but since the behavior of such packets is known to be independent of the light intensity, it is clear that the result would be the same if only one photon were involved. Diffraction will make trouble if we try to define the path too accurately by using an excessively small aperture, but diffraction will be minimized by using x-ray wave-lengths. We will return to this question later on.

Clearly if we are to associate a wave motion with the orbital motion of a large scale body such as a golf ball, the wave function must be of the same nature as that of the optical wave packets just described. Otherwise we could not interpret the observed properties of our golf ball (simultaneous position and velocity) by means of waves. Hence if the analogy between Fermat's principle and the principle of least action really means anything, it must be possible to show from it that the orbit and orbital velocity of a large scale "particle"[5] are identical with the orbit and orbital velocity of a wave packet in a suitably defined wave problem. We must then show (a) that wave packets travel along the rays of geometric optics and (b) that the speed of the packet is the same as the speed of the corresponding mechanical particle.

The first of these two propositions is commonly assumed without proof. Professor Slater has worked out a simple elementary proof (unpublished) for the general case of a non-homogeneous medium, but we shall not reproduce the argument here. Consider then the second of these two propositions. The speed of the packet is the same as the group velocity of the waves, v_g. According to a well-known formula

$$\frac{1}{v_g}=\frac{\partial}{\partial\nu}\left(\frac{1}{\lambda}\right). \tag{8}$$

[5] Of course, a large scale body is an assemblage of many particles. The justification for treating it as a single particle in the wave mechanics follows from a center of gravity theorem as in Newtonian mechanics.

If this is to be the same as the velocity of the particle v, Eq. (7) gives

$$\frac{1}{v}=\frac{\partial}{\partial\nu}\left(\frac{p}{C}\right)=\frac{\partial}{\partial\nu}\left\{\frac{[2\mu(E-V)]^{1/2}}{C}\right\}=\frac{\mu}{Cp}\quad\frac{\partial E}{\partial\nu}. \tag{9}$$

As p is equal to μv, this reduces to

$$C=\partial E/\partial\nu, \tag{10}$$

a relation which immediately suggests the assumption that the Einstein energy-frequency relation (1) holds for matter as well as for radiation. If we accept (1) as an hypothesis for matter, Eq. (10) is satisfied and we see that the required relations between the motion of the particle and the packet will be fulfilled if

$$p\equiv\mu v=[2\mu(E-V)]^{1/2}=h/\lambda. \tag{11}$$

This equation has been confirmed experimentally by the electron diffraction experiments of Davisson and Germer,[6] Thompson[7], and Rupp.[8] The wave-lengths involved are almost inconceivably small when (11) is applied to bodies of microscopic dimensions. In the case of a golf ball weighing 47 grams the ratio h/μ is 1.4×10^{-28}. If the ball moves with a speed as low as a millimeter in 10 seconds, the wave-length is 1.4×10^{-26} centimeters! This means that diffraction effects are hopelessly beyond the reach of experiment in the case of large scale bodies. On the other hand, the computed wave-length is not inappreciable when Eq. (11) is applied to atomic and molecular problems. For example, the wave-length of an oxygen molecule with a speed corresponding to the mean thermal energy at 300°K is approximately 1.45×10^{-8} cm, while that for an electron with a 10-volt kinetic energy is 5.3×10^{-8} cm. As these wave-lengths are of the order of magnitude of atomic diameters, it is clear that diffraction effects must play a prominent part in atomic dynamical problems.

It is important to note that by introducing a vector $\boldsymbol{\sigma}$ having the magnitude $1/\lambda$ and the direction of the wave normal, we can throw equation (11) into the vector form

$$\boldsymbol{p}=h\boldsymbol{\sigma}. \tag{12}$$

We shall call $\boldsymbol{\sigma}$ the vector wave number. Its components σ_x, σ_y, σ_z denote the number of waves per centimeter crossed by lines parallel to the x, y, and z axes.

5. The Schrödinger Wave Equation

In view of the above results we may assume that the differential equation (4) is valid for matter waves if we set

[6] C. Davisson and L. H. Germer, Phys. Rev. **30**, p. 705 (1927); Proc. Nat. Acad. **14**, p. 317 (1928).

[7] G. P. Thomson, Proc. Roy. Soc. **117**(A), p. 600 (1928); **119**(A), p. 651 (1928).

[8] E. Rupp, Ann. d. Physik **85**, p. 981 (1928).

$$w(x,y,z,\nu)=\lambda\nu=\frac{E}{p}=\frac{E}{[2\mu(E-V)]^{1/2}}=\frac{h\nu}{[2\mu(h\nu-V)]^{1/2}}. \tag{13}$$

As in the optical case the differential equation holds for monochromatic or "monoenergetic" wave functions only, it being assumed that all possible wave forms can be built up from monochromatic waves.

This restriction means that all solutions of Eq. (4) are to be of the form

$$\psi=u(x,y,z)T(t) \tag{14}$$

were $T(t)$ is of the form

$$T(t)=\cos 2\pi(\nu t+\epsilon) \text{ or } e^{\pm 2\pi i(\nu t+\epsilon)}.$$

Then substitution from Eqs. (13) and (14) into Eq. (4) and elimination of T yields

$$\nabla^2 u+\frac{8\pi^2\mu}{h^2}(E-V)u=0. \tag{15}$$

This latter equation is quite as general as (4). The factor $u(x, y, z)$ is sometimes called the amplitude. The term "space factor" is also appropriate and will be used in this review.

For many purposes Eqs. (4) and (15) are adequate, but we frequently have to do with problems in which ψ is made up of a sum of terms of the type of Eq. (14) and for which we need a more general differential equation which does not contain the parameter E or its equivalent ν. Such a differential equation becomes a necessity when we have to do with problems in which the potential energy depends on the time, or in which for any reason the energy of the system is not conserved. A most important example is the perturbation of an atom by an external light wave which is the basis of the theory of dispersion. Here the assumption that we have to do with a single monochromatic wave function or a fixed combination of such functions breaks down completely.

The obvious procedure for deriving a differential equation applicable directly to a wave packet, or to the sum of several monochromatic waves is to eliminate ν from Eq. (4) by differentiation. Schrödinger[9] has shown that in this way one obtains the rather awkward fourth order equation

$$\left[\nabla^2-\frac{8\pi^2}{h^2}\mu V\right]^2\psi+\frac{16\pi^2\mu^2}{h^2}\frac{\partial^2\psi}{\partial t^2}=0. \tag{16}$$

This equation is correct for any monochromatic ψ function which obeys Eq. (4) or Eq. (14), or for any linear combination of such functions. A simpler equation of the second order, however, is adequate for our needs. Again following Schrödinger we observe that if we specialize $T(t)$ [Eq. (14)] making the assumption that the basic monochromatic functions have the form

$$\psi=u(x,y,z)e^{+2\pi iEt/h} \tag{17}$$

[9] E. Schrödinger, Ann. d. Physik (4) **81**, p. 109 (1926).

we can replace $E\psi$ by $(h/2\pi i)(\partial\psi/\partial t)$ thus converting Eqs. (4) and (15) into

$$\nabla^2\psi - \frac{8\pi^2\mu}{h^2}V\psi - \frac{4\pi\mu i}{h}\frac{\partial\psi}{\partial t} = 0. \tag{18}$$

This latter equation[9] is much easier to handle than (16) and is equally adequate for the description of the physical facts. The specialization of the function $T(t)$ involved in Eq. (17) is of the nature of a convention. It is justified by its convenience, and by the fact that waves of this type have all the characteristics needed for our physical problem.

Eq. (18) also makes a simple connection with the Hamiltonian function of the classical mechanics which is of great importance when we come to extrapolate our formulas to non-conservative systems where wave functions of the form (17) do not exist.

Of course the convention (17) could not be used for optical problems since the electric and magnetic vectors which obey the optical differential equation have an intrinsic physical meaning which requires them to be real. Our freedom to adopt Eq. (17) comes from the fact that while $|\psi|^2$ has a direct physical meaning, ψ has not.

6. Application of Restricted Relativity Principle to Free Particle

Historically the first step in the development of wave mechanics was the application of the restricted relativity principle to the problem of the correlation of waves and free particles by de Broglie.[10] Symmetry demands that a stationary particle be associated with a stationary wave system. In other words the wave function for a free particle must be of the form

$$\psi = f(x_0, y_0, z_0)e^{2\pi i\nu_0 t} \tag{19}$$

when referred to a system of coordinates x_0, y_0, z_0 with respect to which the particle is at rest. To get the form of the wave for an electron moving in the direction of the z axis de Broglie applies the Lorentz transformation,

$$z_0 = \frac{z - vt}{[1 - v^2/c^2]^{1/2}}; \quad t_0 = \frac{t - vz/c^2}{[1 - v^2/c^2]^{1/2}}; \quad x_0, y_0 = x, y$$

which yields

$$\psi = f\left(x, y, \frac{z - vt}{[1 - v^2/c^2]^{1/2}}\right) \exp\left\{\frac{2\pi\nu_0}{[1 - v^2/c^2]^{1/2}}\left(t - \frac{vz}{c^2}\right)\right\}. \tag{20}$$

This expression may be made to describe either an infinite plane wave system or a wave packet according to the hypothesis regarding the amplitude $f(x_0, y_0, z_0)$. In either case the frequency defined by the phase factor is

[10] L. de Broglie, Nature, **112**, p. 540 (1923); Thesis, Paris (1924); Ann de Physique, (10) **3**, p. 22 (1925).

$$\frac{\nu_0}{[1-v^2/c^2]^{1/2}}.$$

If the zero level of energy is fixed in accordance with the usual relativity expression

$$E=\mu c^2=\frac{\mu_0 c^2}{[1-v^2/c^2]^{1/2}} \tag{21}$$

E and ν transform in the same way so that the fundamental relation (1) is invariant of a Lorentz transformation. This fact is the first great contribution of de Broglie. The wave-length required by Eq. (20) is

$$\lambda=\frac{c^2}{v}\frac{[1-v^2/c^2]^{1/2}}{\nu_0}=\frac{c^2 h}{Ev}=\frac{h}{\mu v},$$

in agreement with Eq. (11). The phase velocity is c^2/v as may be read directly from (20) or deduced from Eq. (1), the above expression for λ, and the relation $w=\lambda\nu$.

The variation in mass with velocity can also be taken into account when the particle moves in a force field of potential energy V. The action integral in this case takes the form

$$S=\int_A^B (F+E-V)dt=\int_A^B p\,ds \tag{22}$$

where $F=\mu_0 c^2[1-v^2/c^2]^{1/2}$ and $p=\mu v=\mu_0 v/[1-v^2/c^2]^{1/2}$. A comparison of (18) and (3) shows that as in the Newtonian theory the rays defined by Fermat's principle and the mechanical orbits agree if $C/\lambda=p$. In this case the expression for the momentum in terms of the energy is

$$p=\frac{1}{c}[(E-V)^2-E_0^2]^{1/2} \tag{23}$$

but as before we find that

$$\frac{\partial p}{\partial \nu}=\frac{\mu}{p}\frac{\partial E}{\partial \nu}=\frac{1}{v}\frac{\partial E}{\partial \nu},$$

so that the group velocity of the waves is equal to the velocity of the particle if we identify C with $\partial E/\partial\nu$ or h.

7. The Problem of n Particles

Eqs. (15) and (18) can be generalized without difficulty to include the case of a system of n particles moving under the influence of conservative forces. It is convenient for this purpose to introduce the idea of a many-dimensional "coordinate space" similar to that used in the classical statistical mechanics. We need a number of dimensions equal to the number of coordinates of the system, i.e. $3n$, rather than twice that number, as in statistical mechanics. Let the Cartesian coordinates of the particles be $x_1, x_2 \cdots x_{3n}$ and let the

corresponding masses be $\mu_1, \mu_2, \cdots \mu_{3n}$ where, of course, the three masses associated with the cordinates of any one individual particle will be the same. Then any point in our coordinate space will represent a possible configuration of the system and the complete motion of the system is represented by the orbit of a representative point in configuration space. We shall show that the motion of this representative point may be identified with the motion of a wave packet derived from a suitably defined wave equation operating in the coordinate space. To this end the principle of least action is conveniently stated in the form

$$\delta S = \int_A^B 2T dt = \delta \int_A^B \sum_{k=1}^{3n} p_k dx_k = 0 \quad (E \text{ unvaried}) \tag{24}$$

where A and B are the initial and final configurations and p_k denotes the momentum $\mu_k \dot{x}_k$ (or $\partial T/\partial \dot{x}_k$) associated with the coordinate x_k. Fermat's principle for waves derived from the $3n$-dimensional equation

$$\sum_{k=1}^{3n} \frac{\partial^2 \psi}{\partial x_k{}^2} = \frac{1}{w^2(x_1 \cdots x_{3n})} \frac{\partial^2 \psi}{\partial t^2} \tag{25}$$

would be

$$\delta \int_A^B \sum_{k=1}^{3n} \frac{\alpha_k dx_k}{\lambda(x_1 \cdots x_{3n})} = 0, \tag{26}$$

where $\lambda = w/v$ and the α's are direction cosines subject to the condition

$$\sum_{k=1}^{3n} \alpha_k{}^2 = 1. \tag{27}$$

In order to throw Eq. (24) into the form (26), it is necessary to introduce slightly different variables, viz.:

$$X_k = \mu_k{}^{1/2} x_k. \qquad k = 1, 2, \cdots, 3n. \tag{28}$$

Then the momentum associated with X_k is

$$P_k = \frac{\partial T}{\partial \dot{X}_k} = \frac{\partial}{\partial \dot{X}_k} \left(\sum_k \frac{1}{2} \dot{X}_k{}^2 \right) = \dot{X}_k, \tag{29}$$

and the principle of least action becomes

$$\delta \int_A^B \sum_k P_k d\dot{X}_k = 0. \tag{30}$$

The energy equation yields

$$\sum_k (P_k)^2 = 2T = 2(E - V) \tag{31}$$

and (30) can be written in the form

$$\delta \int_A^B \sum_k g_k [2(E-V)]^{1/2} dX_k = 0 \tag{32}$$

if we define the g_k's by

$$g_k = P_k / [2(E-V)]^{1/2}. \tag{33}$$

This definition and (31) show that the g's are of the nature of direction cosines being subject to the same condition as the α's, viz.,

$$\sum_{k=1}^{3n} g_k^2 = 1.$$

The g's [cf. (33)] are clearly determined by the path of the representative point of the system in configuration space and are independent of the speed with which the path is traversed.

Equation (32) states the principle of least action in a form entirely analogous to the generalized Fermat's principle (26), and the problem of identifying the solutions of the two problems is solved if we first assume that equations (25) and (26) are valid only if we replace the x_k's by X_k's and then fix the form of the hitherto arbitrary function w to agree with

$$\frac{C\nu}{w} = \frac{C}{\lambda} = [2(E-V)]^{1/2}. \tag{34}$$

As before we may identify $[2(E-V)]^{1/2}$ with the resultant momentum or resultant velocity of the representative point in $X_1, \cdots X_{3n}$ space and can show that this velocity is equal to the group velocity or packet velocity of the waves if $C = \partial E / \partial \nu$. Introducing the assumption that $E = h\nu$, we find that the motion of wave packets in configuration space is in agreement with the Newtonian mechanics if the wave equation has the form

$$\sum_{k=1}^{3n} \frac{\partial^2 \psi}{\partial X_k^2} = \frac{1}{w^2(X_1, \cdots, X_{3n})} \frac{\partial^2 \psi}{\partial t^2} \tag{35}$$

and

$$w^2 = \frac{h^2\nu^2}{2[h\nu - V]} = \frac{E^2}{2[E-V]}. \tag{36}$$

Returning to our original coordinates and adopting the assumption (17), we derive from Eq. (35) the familiar Schrödinger equation

$$\sum_{i=1}^{n} \frac{1}{\mu_i} \nabla_i^2 u + \frac{8\pi^2}{h^2}(E-V)u = 0 \tag{37}$$

where

$$\nabla_i^2 = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}$$

and i is a particle index replacing the cordinate index k of Eq. (35). Eliminating the energy as before we obtain

$$\sum_{i=1}^{n} \frac{1}{\mu_i} \nabla_i^2 \psi - \frac{8\pi^2}{h^2} V\psi - \frac{4\pi i}{h} \frac{\partial \psi}{\partial t} = 0. \tag{38}$$

While the above discussion clearly reveals the possibility of developing a wave mechanics of which the classical mechanics is an appropriate limiting case it brings out one important distinction between radiation waves and matter waves. The latter must be thought of as functions of as many independent variables as there are coordinates in the mechanical system under consideration, whereas the former are functions of only three independent variables. We may say, if we like, that the matter waves are waves in a $3n$-dimensional space, but this is only one way of describing the fact that there are more than three independent variables in Eq. (27) and Eq. (28). The idea of a many-dimensional space is by no means essential to these equations or to their application. The reason for this difference between matter waves and light waves lies in the fact that matter corpuscles exert forces on each other, while photons do not. Consider the two-body problem of astronomy. If we try to describe the orbital motion of the two bodies by two independent wave packets in three-dimensional space, we confront immediate difficulty in that the behavior of each packet is affected by the instantaneous position of the other. Then the index of refraction for the waves of one packet must be a function of the coordinates of the other packet. Thus the waves of each packet must be functions of six independent variables instead of three.

In special cases, however, such as streams of electrons or atomic rays, we have to deal with assemblages of particles under circumstances which permit us to neglect the forces between them. Such streams can be treated like photons by means of three-dimensional waves.

8. The Statistical Interpretation of the Wave Theory of Matter

Let us now return to the three-dimensional waves of Eqs. (4), (15) and (18) and consider their interpretation in somewhat greater detail. Our assumptions up to this point comprise the following:

(a) Matter, like radiation, is both waves and corpuscles.

(b) In the case of a single particle in a conservative force field all the waves are to be built up out of monochromatic solutions of the wave equation (18).

(c) The behavior of large scale bodies as described by the Newtonian mechanics is to be interpreted in wave language by means of wave packets compounded from elementary approximately plane† waves with a variety of frequencies and wave normals.

† That is, plane over the volume of the packet.

(d) The energy of a particle and the frequency of the associated waves are related according to the optical rule

$$E = h\nu. \tag{1}$$

(e) The vector momentum of the particle and the vector wave-number $\mathfrak{d}$ are related according to the rule

$$\boldsymbol{p} = h\mathfrak{d}. \tag{12}$$

(f) The "intensity" of the wave function in any element of volume $dxdydz$ at the time t is to determine the probability that the particle lies in that element just as the intensity of light waves in an element $dxdydz$ determines the probability that any given photon lies in the corresponding element.

This last assumption requires some clarification. In the case of radiation the intensity is given by the sum of the squares of the electric and magnetic vectors. For simplicity we shall regard the wave function ψ for matter as a complex scalar quantity for the present, however, since such a scalar is adequate to describe all that we know of the wave characteristics of matter. The term "intensity" is accordingly identified with $|\psi|^2$ or with $\psi\psi^*$ where ψ^* is the complex conjugate to ψ.

To give precision to the relation between $\psi\psi^*\,dxdydz$ and the probability that the associated particle lies in $dxdydz$ we note that in the case of a complex radiation field the intensity $(\mathfrak{E}^2+\mathfrak{H}^2)$ determines the probable energy density of the photons rather than the "number density." If, however, we have given a wave function ϕ such that $\phi\phi^*\,dxdydz$ measures probable *energy* density, we can always set up a second function ψ which obeys the same differential equation and such that $\psi\psi^*\,dxdydz$ is at least a plausible measure of the desired number density of the particles. To do so we have merely to analyse ϕ into monochromatic constituents, divide the amplitude of each by $(h\nu)^{1/2}$ and sum up again. This statement will be justified in the discussion of the Fourier analysis of wave packets which follows. Since number density rather than energy density is fundamental in the theory of matter waves we assume that the wave functions we are dealing with are so formed that we may identify $\psi\psi^*\,dxdydz$ with the former type of density at least to a constant of proportionality. If we have to do with a single particle we may fix the absolute magnitude of ψ by means of the normalization condition

$$\int\int\int_{\infty} \psi\psi^* dxdydz = 1, \tag{39}$$

in which case $\psi\psi^*\,dxdydz$ gives directly the probability that the particle lies in the volume element $dxdydz$.[11]

[11] It is easily proven by the aid of Green's theorem that if ψ is any solution of Eq. (18) which vanishes at infinity

$$\frac{\partial}{\partial t}\int\int\int_{\infty} \psi\psi^* dxdydz = 0.$$

It is but a short step from these assumptions to the formulation of the Heisenberg uncertainty principle which forbids the assignment of precise simultaneous values to the coordinates of a particle and to the corresponding momenta. In order to develop this principle, we observe first of all that the frequency ν and the vector wave-number $\mathfrak{s}$ are uniquely defined only for infinite plane monochromatic waves. A wave packet, like an experimental spectrum line, contains a narrow continuous spectrum of frequencies and also a continuous "spectrum" of wave normals. Hence the right hand members of Eq. (1) and Eq. (12) are to a certain extent ambiguous. Conceivably we may interpret the symbols ν and $\mathfrak{s}$ as mean values over the packets under consideration, or we may assume that the range of values of ν and $\mathfrak{s}$ contained within a wave packet is to be correlated with a corresponding uncertainty in the energy and momentum of the associated particle. The former assumption, however, is readily shown to be untenable. To prove this let us consider the case of a wave function composed of the sum of two disturbances ψ_1 and ψ_2 each of which is a typical wave packet having a fairly well defined frequency and direction of motion. In the optical case such a disturbance could be obtained experimentally by allowing a beam of plane parallel monochromatic radiation to fall on an aperture covered by a shutter. Opening and closing the shutter momentarily would form a primary packet which could be split into two parts with different wave-lengths and directions of motion by suitable reflection from a moving half-silvered glass mirror. If a disturbance of this kind is associated with a single particle, whether electron or photon, hypothesis (*f*) requires that the particle have a certain probability of moving with packet No. 1 with its corresponding velocity and a certain complementary probability of moving with packet No.2 with its velocity. But these two possible positions are correlated with two different momenta whose probabilities are determined by hypothesis (*f*). Hence in this case the correlation of a single energy and single momentum with the wave function is impossible, and since no dividing line can be drawn between such a case as this and the case of the typical single packet we conclude that it is equally impossible for a single packet. In fact we know by direct experiment in the optical case that such a single packet always contains a range of possible energies and momenta. If the packet be formed with the aid of an aperture and shutter as suggested above, diffraction will cause the radiation to diverge by an amount varying inversely with the dimensions of the aperture. If the intensity is large and the aperture is small compared with the wave-length, photons will proceed in all directions from the slit carrying with them momentum directed along the radius vector from the slit. At the same time the interruption of the primary beam by the shutter will destroy its monochromatic character and scatter the photons over a narrow continuous spectrum. Diffraction experiments involving prolonged photographic exposures and very low intensities show that the distribution of energy over the pattern is independent of the intensity. We infer that if the intensity of the radiation in the packet under consideration is so low that only one or

two photons pass through the aperture during its formation, the relative probability of each energy and direction of motion is the same as if the intensity were very large.

The determination of the relative probabilities of the different momenta associated with a wave packet moving in a field free space is fundamental for our theory. As a first step in the derivation of a formula for this purpose we observe that the usual elementary theory of Fraunhofer diffraction phenomena in which one sums up the contributions of Huygens' wavelets from different parts of the aperture is equivalent to an analysis of the radiation emerging from it into plane waves having different wave normals. At large distances from the aperture where the Fraunhofer theory applies, the energy of the radiation in any elementary cone of solid angle $d\Omega$ is equal to the energy of a corresponding set of plane waves in what we may call the "plane wave spectrum" of the complete wave system at the aperture. Hence to find the probability that any particular photon in a monochromatic diffracted train has a momentum vector lying in the solid angle $d\Omega$ we have to analyse the beam emergent from the aperture into plane waves and compute the ratio of the sum of the intensities of all elementary waves having normals in $d\Omega$ to the intensity of the entire beam. Similarly if we have radiation which is not monochromatic, (whether due to a finite length of train or not) we can find the probability that any photon has an energy in the interval between $h\nu$ and $h(\nu+d\nu)$ by a Fourier analysis of the wave function into rigorously monochromatic constituents. The complete analysis of a finite wave group or packet into a three-dimensional system of plane waves characterized by the three independent parameters σ_x, σ_y, σ_z, will then serve to determine the relative probability of the three components of the momentum and of the different energies of the associated particle or particles.

The formulas for this analysis are simpler for matter waves than for radiation since in the former case the waves are described in terms of complex numbers. The normal form for the description of a plane matter wave is then

$$\psi = C^{2\pi i[\nu t-(x\sigma_x+y\sigma_y+z\sigma_z)]} = Ce^{(2\pi i/h)[Et-(xp_x+yp_y+zp_z)]} \tag{40}$$

where C is a complex constant. This expression is a solution of the wave equation (18) for a homogeneous medium (zero force field) if the frequency satisfies the relation

$$\frac{\nu^2}{w^2} = \frac{1}{\lambda^2} = \sigma_x^2+\sigma_y^2+\sigma_z^2 = h^{-2}(p_x^2+p_y^2+p_z^2). \tag{41}$$

The desired analysis is accomplished by the aid of Fourier's integral which permits us to resolve any wave packet into exponential components by the use of the formulas

$$\psi(t,x,y,z) = \frac{1}{h^{3/2}}\int\int\int_{\infty} Q(t,p_x,p_y,p_z)e^{-(2\pi i/h)(xp_x+yp_y+zp_z)}dp_x dp_y dp_z, \tag{42}$$

$$Q(t,p_x,p_y,p_z) = \frac{1}{h^{3/2}}\int\int\int_{\infty} \psi(t,x,y,z)e^{+(2\pi i/h)(xp_x+yp_y+zp_z)}dx dy dz. \tag{43}$$

The packet may be defined either by ψ or by Q. If the latter function has the form

$$Q(t, p_x, p_y, p_z) = G(p_x, p_y, p_z)e^{-2\pi i\nu t} \tag{44}$$

where ν is defined by Eq. (41), the expression (42) for ψ becomes a solution of the wave equation (18) for the case of a free particle where $V=0$. It can be fitted to arbitrary initial conditions provided only that at $t=0$ ψ vanishes at infinity, and so will represent any wave packet. Moreover, it may be proved that[12]

$$\iiint_{\infty} \psi\psi^* dxdydz = \iiint_{\infty} QQ^* dp_x dp_y dp_z \tag{45}$$

and we conclude that as $\psi\psi^*\ dxdydz$ measures the probability that the particle lies in the volume element $dxdydz$ in ordinary space, so $QQ^*\ dp_x\ dp_y\ dp_z$, or $GG^*\ dp_x\ dp_y\ dp_z$, measures the probability that its momentum vector terminates in the element $dp_x\ dp_y\ dp_z$ of "momentum space."[13] Q describes a stationary wave motion in momentum space and satisfies a differential equation that is closely related to that of ψ. Jordan calls $\psi(t, x, y, z)$ the *probability amplitude* for the positional coordinates and $Q(t, p_x, p_y, p_z)$the probability amplitude for the momentum.

In case we have to do with the motion of a particle in a force field the momentum varies with t and the plane wave (40) is not a solution of the wave equation, though it may approximate to a solution in a suitable small neighborhood. Hence $\psi(t, x, y, z)$ as defined by Eq. (42) does not describe the movement of the packet in time if p_x, p_y, p_z are treated as independent of t, but we can still use equations (41) and (43) to determine the instantaneous probability distribution for the momentum components of the particle. This extension of equations (42) and (43) to the more general case is justified by our initial correlation of wave-length and momentum through the principles of least time and least action.

[12] This is a limiting case (in three dimensions) of the well-known theorem that if

$$f(x) = \sum_{\tau} c_\tau e^{2\pi i\tau/T}, \qquad a<x<a+T,$$

$$\sum_{\tau} c_\tau c_\tau^* = \int_a^{a+T} f(x)f^*(x)dx$$

Cf. Schuster, Phil. Mag. **37,** p. 509 (1894); Schuster and Nicholson, Theory of Optics, 3rd Ed. 1924, p. 334.

[13] If $QQ^*dp_xdp_ydp_z$ gives the probability that the momentum vector terminates in the element $dp_xdp_ydp_z$, it is clear that $h\nu QQ^*dp_xdp_ydp_z$ gives the energy associated with that element. Then if $\phi(t, x, y, z)$ is defined by

$$\phi = \frac{1}{h^{3/2}} \iiint_{\infty} (h\nu)^{1/2} Q \exp\left[-\frac{2\pi i}{h}(xp_x+yp_y+zp_z)\right] dp_x dp_y dp_z$$

$\phi\phi^*dxdydz$ must give the energy associated with the volume element $dxdydz$ in ordinary space. Conversely, given ϕ we can pass to ψ as stated on p. 170.

9. Heisenberg's Uncertainty Principle[14]

The qualitative discussion in the preceding article shows clearly that a reduction in the size of a wave packet to *increase* the precision with which it locates the associated particle in space is always accompanied by a *decrease* in the precision with which its momentum is defined. The question immediately arises "Is the impossibility of forming a wave packet which will describe a particle with a perfectly definite position and momentum a weakness of the wave mechanics or does it correspond to an inherent experimental difficulty in making precise simultaneous observations of position and momentum?" Heisenberg has shown that the second of these alternatives is correct and on this basis formulates the principle: *Under ideal conditions the product of the experimental uncertainty in the value of any coordinate q and the experimental uncertainty in the value of the corresponding momentum p has a minimum value of the order of magnitude of Plank's constant h.*[15]

If we grant for the moment that the minimum experimental uncertainties are those inherent in our wave packet theory we may deduce Heisenberg's principle from equations (42) and (43). Consider the case of a packet bounded at $t=0$ by a rectangular box such that in the region S defined by the inequalities $-a<x<+a$; $-b<y<+b$; $-c<z<+c$, ψ_0 has the value

$$\psi_0 = A e^{-(2\pi i/h)xp_0}$$

whereas outside S, ψ_0 vanishes. In this case we may identify the uncertainties in x, y, and z with a, b, and c respectively. The momentum of the packet is as sharply defined as is possible for a packet of this size, the mean values of the components being p_0, 0, 0. Applying Eq. (42) we have

$$Q(0, p_x, p_y, p_z) = G(p_x, p_y, p_z) = \frac{8abc}{h^{3/2}} G_1(p_x) G_2(p_y) G_3(p_z)$$

where

$$G_1(p_x) = \frac{1}{2a}\int_{-a}^{+a} e^{-(2\pi i/h)x(p_0-p_x)}dx = \frac{\sin\{2\pi a(p_0-p_x)/h\}}{2\pi a(p_0-p_x)/h},$$

$$G_2(p_y) = \frac{\sin 2\pi b p_y/h}{2\pi b p_y/h}\ ; \qquad G_3(p_z) = \frac{\sin 2\pi c p_z/h}{2\pi c p_z/h}$$

($|A|$ takes on the value $(8\ abc)^{1/2}$ if adjusted to satisfy the normalizing condition Eq. (39)). The probability that p_x lies in the interval dp_x is then proportional to

$$G_1^2(p_x)dp_x = \frac{\sin^2 2\pi a(p_0-p_x)/h}{[2\pi a(p_0-p_x)/h]^2}\,.$$

[14] W. Heisenberg, Zeits. f. Physik **43**, 172 (1927); N. Bohr, Nature **121**, 580 (1928).

[15] The uncertainty postulated is, of course, quite unimportant for the dynamics of large scale bodies. If the uncertainty in any one of the positional coordinates of the golf ball of p. 163 is of the order of magnitude of an atomic diameter the corresponding minimum uncertainty in velocity is of the order of magnitude of 10^{-20} cm/sec!

This distribution function is the familiar one met with in the theory of diffraction by a single slit and illustrated by Fig. 1. The half spacing of the central pair of minima ($\xi = \pm\pi$ in Fig. 1) gives a fair measure of the uncertainty in the value of p_x. Hence $\delta p_x = h/2a$ and

$$\delta x \delta p_x (= \delta y \delta p_y = \delta z \delta p_z) = h/2 \qquad (46)$$

in agreement with Heisenberg's principle.†

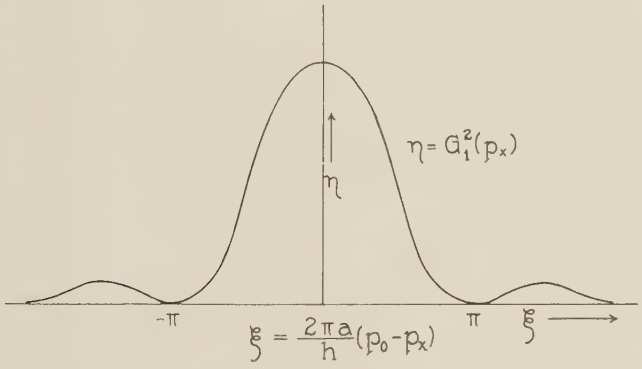


Fig. 1

The same type of formula may be used to relate the uncertainty in the energy of the particle and the uncertainty in the time at which it takes on a certain coordinate value. For example, the above wave packet will cross the plane $x = 0$ in the time $2\delta t$ given by

$$2\delta t = 2a/v = 2a\frac{\partial \sigma_x}{\partial \nu} = \frac{2a}{h}\frac{\partial p_x}{\partial \nu}.$$

Hence the product of the uncertainties in t and E is

$$\delta E \delta t = h\delta\nu \frac{a}{h}\frac{\partial p_x}{\partial \nu} = a\delta p_x = h/2.$$

In the Hamiltonian theory of classical dynamics t and $-E$ are canonically conjugate variables like x and p_x. Hence a plausible extrapolation would lead us to expect the same relation between the uncertainties in any two canonically conjugate dynamical variables.

Heisenberg justifies the conclusion that these uncertainty relations are experimental as well as theoretical by an analysis of the various possible experimental means for determining simultaneous values of coordinate and momentum. We may determine the position of a particle, for example, by observing the direction of motion of photons or electrons which have been scattered by collision with it or by allowing it to pass through a slit or small aperture. In each case, however, if one assumes the validity of the ordinary theory of the Compton effect and also the applicability to matter

† In the one-dimensional case if the packet is defined to give $\psi\psi^*$ the form of a Gauss error curve GG^* has the same form and $\delta x \delta p_x = h/2\pi$. Cf. W. Heisenberg, reference 14.

of the optical theory of diffraction by apertures, one finds that the observation of any positional coordinate introduces an uncontrollable alteration in the corresponding component of momentum such that even if the latter were uniquely known before the position was observed, Heisenberg's uncertainty relation would hold true after the observation of position. Conversely if one measures the positional coordinate first and the momentum afterward, the measurement of momentum introduces an uncontrollable alteration in position with the same ultimate result regarding the minimum values of the uncertainties.[16] This savors of arguing in a circle since we assume the diffraction of matter in order to derive from a conceptual experiment a relation previously obtained directly from the theory of diffraction. The point is, however, that there is no way of sidestepping the diffraction effect if it is real—a qualification which need hardly be considered.

As Bridgman[17] has shown, the uncertainty relation is bound up with our inability to trace out the details of collisions between photons, electrons, and apertures. This inability, in turn, is due to the absence of tools finer than complete collisions for making the measurements necessary to give reality to such details.

The perfect harmony which exists between our dualistic theory and the experimental limit of the uncertainty product $\delta q \delta p$ is a strong indication that the former is on the right track. It shows that the theory is marvelously adapted to the description of observable magnitudes without introducing others which are unobservable as did the Bohr theory. The desirability of eliminating unobservable quantities from the theory was in fact one of the basic ideas in Heisenberg's original development of the matrix mechanics.

It is hardly necessary to dwell here on the philosophical significance of the uncertainty principle which gives final precise form to the hypothesis of indeterminism in physics. This hypothesis may be said to have originated in the discovery of the law of radioactive decay. It received important support in Einstein's speculations on the transition probabilities which govern the jumps of atoms from one energy level to another and now seems permanently enthroned in the quantum mechanics. To be sure, many physicists still regard the indeterminism of the quantum theory as a temporary phase in the development of the science and look for a return to determinism later on. To the writer, however, it seems quite clear that such a return is impossible unless some wholly revolutionary discoveries of an experimental character give us tools which we now lack for the detailed investigation of collisions. Moreover, the quantum mechanics has so many of the characteristics of ultimate perfection in theory that the possibility of unlimited revolutions in the future can no longer be granted without question. If the

[16] For a detailed discussion of the various possible types of experiment the reader is referred to the original papers of Heisenberg and Bohr (note 12). It is perhaps desirable to note here that we always observe pairs of coordinates and not single coordinates. Thus the "simultaneous" measurement of a coordinate q and conjugate momentum p means the correlation of the values of q and p with the same value of some third variable x, which may be either some independent space coordinate or the time.

[17] P. W. Bridgman, Harper's Magazine, March, 1929, p. 443.

purpose of science is to reveal the orderliness and inner simplicity of nature, scientific theory is to be judged by the extent to which it reduces chaos to a single formula or group of formulas. If we look back over the history of science and forward to the future with this criterion in mind, we can hardly resist the conviction that we are within sight of the goal. The past has seen the reduction of existence to the two modes, matter and radiation, the reduction of matter to the 92 chemical elements and their isotopes, the reduction of the various atomic species to electrons and protons, and most recently the lowering of the barrier between matter and radiation through the discovery of their common dualistic nature and of the possiblity of transmuting one into the other.[18] Many problems have yet to be solved before the territory already within our grasp shall have been fully consolidated, but in view of the uncanny power of the tools now available the discrepancy between the present state of physical theory and the ideal is almost uncomfortably small.

SECTION 2. THE CHARACTERISTIC VALUE PROBLEM OF THE WAVE MECHANICS

1. The Linear Oscillator

It is a well-known empirical fact that each species of atom has a characteristic set of discrete energy levels. From the wave mechanics point of view these levels are to be identified with a corresponding set of standing wave monochromatic solutions of the equation (38) for the atom in question. In order to obtain such a set of solutions Schrödinger has introduced the hypothesis[19] that the only solutions of the wave equation having physical meaning are those which are finite, single-valued and twice differentiable over the whole of coordinate space.[20] The hypothesis is justified by its results and by the fact that infinite or multiple-valued functions would be hard to interpret physically. It is equivalent to a boundary condition (and will be referred to hereafter as such) since in many cases ψ must either become infinite at singular points of the differential equation or satisfy typical homogeneous boundary conditions there.

The problem thus presented is of the familiar type known to the mathematicians as the characteristic value problem (Eigenwertproblem).[21] Solutions are obtainable only for certain characteristic values (Eigenwerte) of an adjustable parameter involved in the differential equation—in this case the parameter E or its equivalent ν. The solutions of the equation which satisfy the boundary conditions are called characteristic functions (Eigen-

[18] The evidence for transmutation is primarily astronomical.

[19] E. Schrödinger, Ann. d. Physik (4) **79,** p. 361 (1927).

[20] It now seems necessary to modify the first part of this hypothesis in certain cases to permit the wave function to become infinite at isolated points provided that $\int\psi\psi^* d\tau$ is finite when extended over small regions surrounding and including these points. Cf. A. H. Wilson, Proc. Roy. Soc. **A118,** p. 635 (1928).

[21] For a comprehensive mathematical discussion the reader is referred to Courant-Hilbert "Methoden der Mathematischen Physik I," Berlin 1924, or to Riemann-Webers "Differentialgleichungen der Physik I," 7th ed. Braunschweig, 1925.

funktionen). The normal modes of vibration of various elastic systems are familiar examples of such characteristic functions.

As a first example of the characteristic value problem of the wave mechanics let us consider the case of a particle of mass μ vibrating in one dimension under the influence of a potential energy function $V(x)$ which for simplicity we assume to have the form shown in Figure 2 with a single minimum at x_0, a pole of the second or higher order at $x=0$, and an asymptotic finite limit at $x=\infty$.[22] (The potential energy V is defined only for the range of values $0<x<\infty$ since the discontinuity in V at $x=0$ would in any

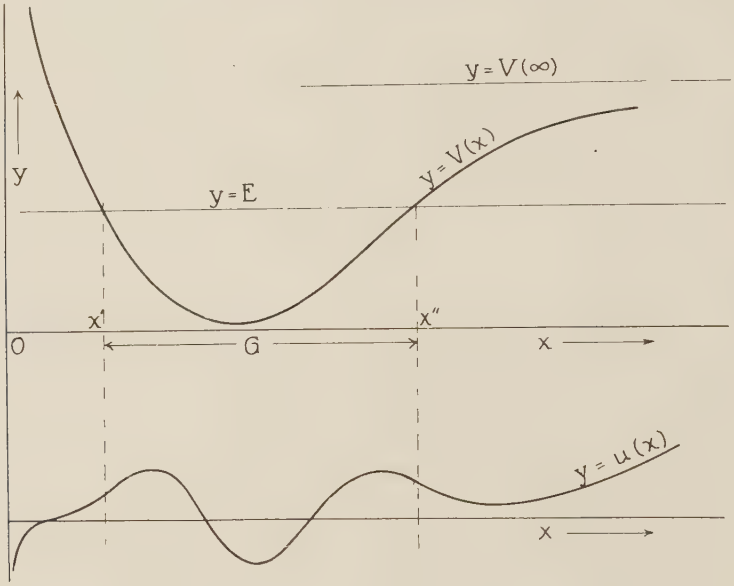


Fig. 2. Potential energy function for one-dimensional oscillator and typical integral curve for Eq. (48).

case destroy the continuity of the solutions $u(x)$ at that point. In the general case of a differential equation to be solved subject to boundary conditions the region defined by the boundaries is called the *fundamental region*. In the applications to wave mechanics values of the independent variable outside the fundamental region have no physical meaning). The wave-length is

$$\lambda(x)=\frac{w(x)}{\nu}=\frac{h}{[2\mu(E-V)]^{1/2}} \tag{47}$$

and the wave equation for the space factor u reduces to the form

$$\frac{d^2u}{dx^2}+\frac{8\pi^2\mu}{h^2}(E-V)u=0. \tag{48}$$

[22] This is the characteristic type of potential energy function for the problem of the vibrations of a diatomic molecule.

λ has a minimum value at x_0 and if $E < V(\infty)$ λ becomes infinite at two points x' and x'' where $E = V$. The classical motion is confined to the region G between x' and x'' if $E < V(\infty)$ (Case I) and extending from x' to infinity if $E \geqq V(\infty)$ (Case II). Outside this region the kinetic energy, defined in the wave mechanics by

$$T = E - V = -\left(\frac{h^2}{8\pi^2\mu}\right)\frac{1}{\psi}\frac{d^2\psi}{dx^2},$$

is negative, while the momentum and wave-length are imaginary. Such negative values of the kinetic energy have no meaning in the classical mechanics but in the present theory they play an important rôle. The wave functions of the linear oscillator problem under discussion are spread out over the whole of the positive x axis including the part outside of G. Since $\psi\psi^* dx$ $(= uu^* dx)$ gives the probability that the particle is in the elemental region dx, it is clear that it must spend a definite fraction of its time in the region of negative kinetic energy. This fraction is small, however, especially for the higher modes of vibration or upper energy levels.

2. Graphical Discussion of the Integral Curves[23]

As the coefficients of Eq. (48) are real, the real and imaginary parts of $u(x)$ must be solutions of the equation which satisfy the boundary conditions. Real solutions are readily shown to be unique except for a constant multiplicative factor and hence any characteristic function $u(x)$ may be resolved into the product of a real function of x and a complex constant. Obviously it suffices to search for real solutions. Each will be characterized by two constants of integration α and β of which one (say α) is multiplicative. An "integral curve" representing a solution of the equation will pass through any predetermined point of the u, x plane with any predetermined slope, but in general these curves will not fit the boundary conditions. Clearly each curve will be concave to the x axis if

$$\frac{1}{u}\frac{d^2u}{dx^2} \quad \text{or} \quad \frac{8\pi^2\mu}{h^2}(V - E)$$

is negative and convex to the axis if it is positive. Consequently each curve (Cf. Fig. 3) is of an oscillatory character like a distorted sine curve inside G, while outside G it is convex to the axis and acts as if repelled from the axis by a force proportional to $V - E$ and to its own absolute value. x' and x'' are points of inflection.

It is easy to show that if V becomes infinite at $x = 0$ as $1/x^2$, or more rapidly, as we shall suppose, $u(x)$ can remain finite at the origin only by vanishing there. Similarly in Case I, where $E < V(\infty)$, $u(x)$ must vanish at $x = \infty$ in order to remain finite there. Moreover, as u is convex to the axis outside of G, it cannot have more than one node on either side of G. The boundary conditions are satisfied and u becomes a characteristic function

[23] Cf. F. Hund, Zeits. f. Physik **40**, 742 (1927).

if the two possible nodes outside G occur at the limiting points $x=0$ and $x=\infty$.

Consider now the adjustment of the parameters α, β, E which must be made in order to locate a characteristic function and characteristic value. The multiplicative parameter α is useless since a variation in its value does not affect the "zero's" of the function u. The condition at the origin can be met by the adjustment of β only. In Case II this is the only boundary condition and every value of E is a characteristic value. In this case the classical motion of the particle would be aperiodic and the continuous "spectrum" of energy values permitted by the wave mechanics corresponds to the continuous spectrum of E values permitted for an aperiodic motion by the Bohr theory. On the other hand in Case I the classical motion is periodic and in the wave mechanics the additional boundary condition at $x=\infty$ can be met only by a suitable adjustment of E.

It is of interest to note the exact way in which the variation in E affects the integral curves and makes it possible to meet the boundary conditions. Starting with any initial value, let E be increased by a small amount ΔE. Let β be adjusted for each value of E to fit the boundary condition $u=0$ at the origin. The increase in E makes the curves less convex to the axis outside G and more concave to the axis inside G. Hence it may be proved either graphically or analytically that all the nodes are shifted to the left.[24] As E increases continuously new nodes appear at $x=\infty$, then move to the left, and finally enter the region G where they accumulate. The first appearance of each node at $x=\infty$ marks a corresponding characteristic function and characteristic value of E. The minimum number of nodes between the boundary points is zero, and the characteristic function having no nodes except those at the boundary points is associated with the lowest characteristic value of E. Two characteristic functions having the same value of E must have the same zeros and can differ only in the multiplicative constant α. For our present purpose they may be identified with one another. Thus we have the rule that the nth characteristic function, or wave function, divides the region between the boundary points into exactly n parts. The number of nodes plays the rôle of the quantum number in the Bohr theory. It is an integer which defines the characteristic function and characteristic value, giving the ordinal number of the latter in a series arranged according to magnitude. Conventional practise counts only the nodes between the boundary points so that the quantum number for the lowest state is zero.[25]

Figure 3 shows the qualitative form of the characteristic functions u for the lowest energy levels.

The general appearance of these lowest wave functions is unaffected by considerable changes in $V(x)$. Suppose, for example, that V remains finite from $x=+\infty$ to $x=-\infty$, with a single minimum between. As before there will be a continuous spectrum of E values for Case II and a discrete

[24] Cf. Courant-Hilbert, p. 367 or Riemann-Weber, p. 281.

[25] Half-integral quantum numbers do not appear except in connection with the electron spin which we ignore for the present.

spectrum for Case I. The lower wave functions for Case I will be qualitatively unchanged in character. If V becomes infinite at $x = \pm\infty$, as in the case of the ideal linear oscillator where $V = 1/2\ kx^2$, the continuous spectrum disappears, but the lower wave functions are substantially unchanged. This remark holds good also if V has two poles of the second or higher order at finite points x_0 and x_1 with a single minimum between them.

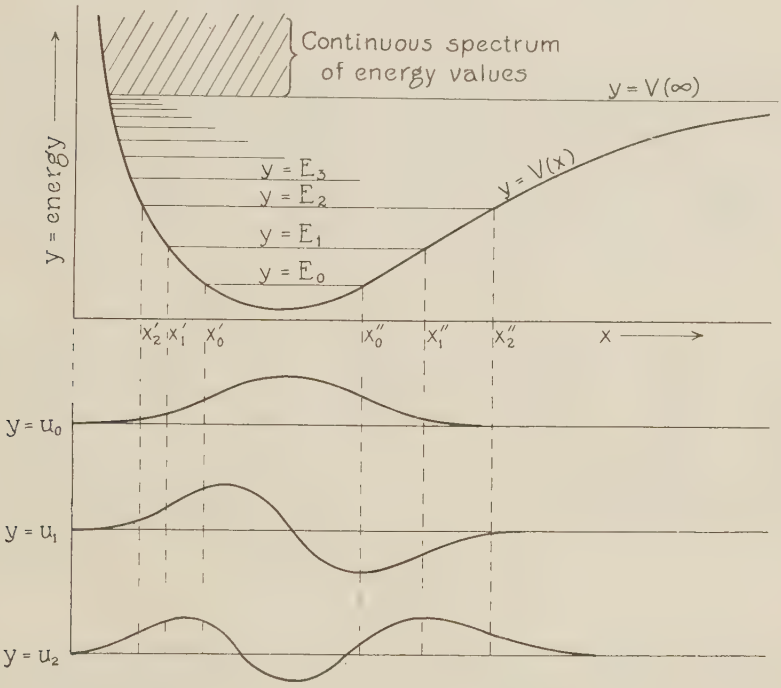


Fig. 3. Energies and wave functions for one-dimensional oscillator.

3. The Sturm-Liouville Problem

All the above mentioned variations in the linear oscillator problem and also most if not all of the exactly solved characteristic value problems of the wave mechanics are ultimately referable to one-dimensional equations of the Sturm-Liouville type

$$\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] - q(x)y + \lambda\rho(x)y = 0 \tag{49}$$

in which λ is the variable parameter playing the rôle of E in Eq. (48).[25a] Solutions of this equation subject to the homogeneous boundary conditions

$$y(a) + \gamma y'(a) = 0\ ; \qquad y(b) + \eta y'(b) = 0 \tag{50}$$

have been studied in detail and shown to have the characteristics described

[25a] In Eq. (49) we have adhered to the notation of Courant-Hilbert despite the fact that p and λ are used in a different sense throughout most of this review.

in the preceding paragraph.[26] (It is possible to show that if Schrödinger's requirement that the wave functions remain finite at the boundary points leads to any boundary condition at all, it will lead to one of the type (50).) The characteristic functions also satisfy an "orthogonality" condition to which we will return later on.

4. Correlation of Characteristic Values with Energy Levels of the Bohr Theory

The possibility of deriving the experimental energy levels of atoms from the wave theory was first indicated by de Broglie[27] who showed that the Bohr quantum condition for circular orbits in hydrogen is identical with the condition that the "optical path" around the orbit is an integral number of wave-lengths. Later Schrödinger[28] showed that in a number of important special cases the energy values given by appropriate solutions of the wave equation are in substantial agreement with the Bohr theory and with experiment.

A more general proof of the agreement between the Bohr theory and the wave mechanics is due to the work of Brillouin, Wentzel, and Kramers.[29] The oscillatory character of the wave functions in the range of the classical vibration G suggests the possibility of describing them approximately by means of a cosine function with a variable amplitude and wave-length. This type of approximation is of no value outside the range G but that fact does not destroy its usefulness. Substitution shows that the formula

$$u = A\lambda^{1/2} \cos\left\{2\pi \int \frac{dx}{\lambda}\right\}. \tag{51}$$

[λ is defined by Eq. (47)] describes the integral curves very well at points not too close to the boundaries of the range G. At each of the points x' and x'' the oscillatory portion of each of the characteristic functions must make a smooth junction with the exponential-like curve outside of G which then approaches the axis monatonically as one moves away from G. Thus the boundary conditions determine the phase angle of the cosine function in Eq. (51) at the points x' and x''. The exact phase will vary with $V(x)$ but Kramers has shown that normally we obtain a good approximation if we give it the values $-\pi/4$ or $-5\pi/4$ at x' and $\pi/4$ or $5\pi/4$ at x''. Hence the phase difference between the points x' and x'' is an odd multiple of $\pi/2$ or

$$2\pi \int_{x'}^{x''} \frac{dx}{\lambda} = \frac{2\pi}{h} \int_{x'}^{x''} p\,dx = (2n+1)\frac{\pi}{2}, \quad n = 0, 1, 2, \cdots \tag{52}$$

[26] Courant-Hilbert, Kap. VI, esp. pp. 366 7; Riemann-Weber, Kap. VII.

[27] L. de Broglie, Thesis, Chap. III; J. de Physique **7**, p. 327 (1926). Cf., however, M. Brillouin, Comptes Rendus **168**, p. 1318 (1919); **169**, p. 48 (1919); **171**, p. 1000 (1920); J. de Phys. **3**, p. 65 (1922).

[28] E. Schrödinger, l. c. and Ann. d. Physik (4) **79**, 489 (1929).

[29] L. Brillouin, Comptes Rendus **183**, p. 24 (1926); J. de Physique **7**, p. 353 (1926); G. Wentzel, Zeits. f. Physik **38**, p. 518 (1926); H. A. Kramers, Zeits. f. Physik **39**, p. 828 (1926).

which is the Wilson-Sommerfeld quantum condition for the problem with the quantum number made a "half-integer." As this quantum condition with half-integral quantum numbers is known to give a satisfactory expression for the energy of a vibrating molecule, we conclude that the characteristic values of E given by the wave equation may be identified with the corresponding experimental energy values.

5. The Planck Ideal Linear Oscillator

The characteristic values and characteristic functions have been worked out exactly for a number of special cases of the linear oscillator problem among which we may mention here the Planck case where $V=\frac{1}{2}kx^2$. The corresponding wave equation and boundary value problem were well known to mathematicians before the advent of the wave mechanics.[30] The characteristic functions are the Hermitian orthogonal functions

$$u_n = A_n e^{-\alpha x^2/2} H_n(\alpha^{1/2}x) \tag{53}$$

where A_n is an arbitary constant amplitude factor, α is the constant $(2\pi/h)(k\mu)^{1/2}$ and H_n is the nth Hermitian polynomial

$$H_n(\xi) = (2\xi)^n - \frac{n(n-1)}{1!}(2\xi)^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2!}(2\xi)^{n-4} - + \cdots \tag{54}$$

If we determine A_n in accordance with the normalization condition (39) it takes the value $2^{-n/2}(n!)^{-1/2}$. The energy levels are identical with those of the Bohr theory using half-integral quantum numbers, viz.,

$$E_n = (n+\tfrac{1}{2})h\nu_0 \qquad n=0,1,2,\cdots \tag{55}$$

ν_0 denoting the classical vibration frequency $(1/2\pi)(k/\mu)^{1/2}$. The wave functions are alternately even and odd as may be proved directly from the differential equation.

The polynomial method described by Sommerfeld[31] gives a valuable direct mode of attack on this and other exactly solvable Sturm-Liouville characteristic value problems. It does not work in all cases but has a wide range of applicability.

6. The Two Body Problem

This is one of the most important characteristic value problems of the theory. We assume that the potential energy V depends only on the distance r between the particles. Let M denote the total mass $\mu_1+\mu_2$ and let μ be a mass coefficient defined by the relation

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2}. \tag{56}$$

Let ξ, η, ζ denote the coordinates of the center of gravity and let x, y, z

[30] E. Schrödinger, Ann. d. Physik (4) **79**, p. 489 (1926); Courant-Hilbert, p. 261.

[31] A. Sommerfeld, "Atombau und Spektrallinien Wellenmechanischer Ergänzungsband," Braunschweig, 1929, Kap. I, 2.

denote the projections of r along the three axes. Then the space factor equation (37) takes the form

$$\frac{1}{M}\left(\frac{\partial^2}{\partial\xi^2}+\frac{\partial^2}{\partial\eta^2}+\frac{\partial^2}{\partial\zeta^2}\right)u+\frac{1}{\mu}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)u+\frac{8\pi^2}{h^2}(E-V)u=0. \tag{57}$$

Let us seek particular solutions of the form

$$u=u_1(\xi,\eta,\zeta)u_2(x,y,z). \tag{58}$$

Substitution of this expression into Eq. (57) breaks the equation into the two parts

$$\left(\frac{\partial^2}{\partial\xi^2}+\frac{\partial^2}{\partial\eta^2}+\frac{\partial^2}{\partial\zeta^2}\right)u_1+\frac{8\pi^2M}{h^2}E_1u_1=0, \tag{59}$$

and

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)u_2+\frac{8\pi^2\mu}{h^2}(E_2-V)u_2=0 \tag{60}$$

where $E_1+E_2=E$. Clearly E_1 is the energy of the translational motion of the center of gravity and E_2 is the internal energy of the system. Equation (59) is the amplitude equation for a free particle. Its solution is a plane wave or superposition of plane waves. A packet may be built up from such solutions to locate the center of gravity in space if desired.

In order to solve the wave equation for the internal motion one may make another change of variables introducing the spherical coordinates r, θ, ϕ. We repeat the process of splitting the differential equation or "separating the variables" by seeking for particular solutions of the type

$$u_2=R(r)Y(\theta,\phi). \tag{61}$$

Eq. (60) then breaks into two parts with the separation constant α, viz.,

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\frac{8\pi^2\mu}{h^2}\left[E_2-V(r)+\frac{\alpha h^2}{8\pi^2\mu r^2}\right]R=0, \tag{62}$$

$$\frac{1}{\sin^2\theta}\frac{\partial^2Y}{\partial\phi^2}+\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left[\sin\theta\frac{\partial Y}{\partial\theta}\right]-\alpha Y=0. \tag{63}$$

The latter equation is familiar from potential theory[32] where it must be solved in such a way that Y is single valued and continuous over the entire sphere. The variables in (63) may be separated in turn and a solution of the problem is then

$$Y=\Phi(\phi)\Theta(\theta)\times\text{const.} \tag{64}$$

where the values of Φ and Θ are

$$\Phi_m(\phi)=e^{im\phi} \qquad m=0,\pm1,\pm2,\cdots \tag{65}[33]$$

$$\Theta_{l,m}(\theta)=P_l^{|m|}(\cos\theta) \qquad l=|m|,|m|+1,\cdots. \tag{66}$$

[32] Courant-Hilbert, p. 265, p. 420.

[33] The functions $\sin m\phi$, $\cos m\phi$ ($m=0, 1, 2\cdots$) are equally valid.

Here $P_l^k(x)$ is the associated Legendre spherical harmonic

$$P_l^k(x) = (1-x^2)^{k/2} \frac{d^k}{dx^k} P_l(x), \tag{67}$$

$P_l(x)$ being the Legendre polynomial of degree l. α has the characteristic values

$$\alpha = -l(l+1) \qquad l = 0, 1, 2, 3, \cdots . \tag{68}$$

Inserting this expression for α into Eq. (62), dropping the subscript in the symbol E_2, and changing the dependent variable from R to $\Re = rR$, we obtain

$$\frac{d^2\Re}{dr^2} + \frac{8\pi^2\mu}{h^2}\left[E - V(r) - \frac{l(l+1)h^2}{8\pi^2\mu r^2}\right]\Re = 0, \tag{69}$$

an equation identical in form with Eq. (48). Comparing this with the energy equation of the classical theory for the radial motion

$$\left(\frac{dr}{dt}\right)^2 + 2\mu\left[E - V(r) - \frac{M^2}{2\mu r^2}\right] = 0 \tag{70}$$

where M is the angular momentum, we see that in the wave mechanics $(h/2\pi)[l(l+1)]^{1/2}$ plays the rôle of angular momentum. This interpretation will be justified later from another point of view.

The solution of Eq. (69) depends on the form assigned to $V(r)$. In the special case of the atomic Kepler problem $V = -Ze^2/r$ where Z is the atomic number of the nucleus. V vanishes at ∞ and if E is negative (Case I, p. 35) there is a discrete set of energy values given by the familiar Bohr formula

$$E_n = -\frac{2\pi^2\mu e^4 Z^2}{h^2 n^2}. \tag{71}$$

Here $n-1$ is the sum of the nodes in $R(r)$, $\Theta(\theta)$, and the real part of $\Phi(\phi)$. If E is positive (Case II) we have a continuous spectrum of E values.

The wave functions R may be expressed in terms of a complex integral in both cases.[19] In Case I they may also be written in the form

$$R_{n,l}(r) = \rho^l L_{n+l}^{(2l+1)}(\rho) e^{-\rho/2} \times \text{const.} \tag{72}$$

where

$$\rho = \left(\frac{8\pi^2\mu e^2 Z}{h^2 n}\right) r = \frac{2Z}{na_1} r \tag{73}$$

and $L_k^{(\tau)}(x)$ is the τ^{th} derivative of the Laguerre polynomial of degree k. a_1 is the radius of the innermost Bohr orbit for hydrogen.

In contrast to the linear oscillator problem we have here a multiplicity of characteristic functions for each energy level paralleling the multiplicity of orbital types which exist for each level in the Bohr theory. This multiplicity is called *degeneracy* and is of a two-fold character. Due to the spherical sym-

metry of the force field each of the characteristic values of α (i.e., each value of l) is associated with $2l+1$ wave functions of the type of Eq. (64) and characterized by the values $0, \pm 1, \cdots, \pm l$ for the quantum number m. This degeneracy is destroyed by the application of an external magnetic or electric field to the atom. In the former case each of the wave functions given by Eqs. (64), (65), and (66) is associated with a separate energy level. Hence m is frequently called the magnetic quantum number. In the absence of disturbing influences the most general solution of Eq. (63) for any given characteristic value α_l is

$$Y_l = \sum_m C_{l,m}\Theta_{l,m}\Phi_m.$$

Such a function is called a tesseral harmonic.

A more special type of degeneracy is exhibited by the radial Eq. (62) in the special case of a Coulomb force field. The energy depends only on the "total quantum number" n and is independent of the "azimuthal quantum number" l. As l takes on all values between 0 and $n-1$ there are in all $\sum_{l=0}^{n-1}(2l+1)$ or n^2 different sets of values of l and m for each energy level. The most general possible wave function for the energy level E_n may be proved to be an arbitrary linear combination of terms of the type $R_{n,l}\Theta_{l,m}\Phi_m$.

7. Orthogonality and the Expansion Problem

The last statement of the preceding article leads us to the question of orthogonality and series expansion which is one of the most important aspects of the characteristic value problem.

Consider first the case of a one-dimensional motion where the differential equation for the space factor $u(x)$ has the form of Eq. (49). Let us assume that there is no continuous spectrum of energy values. A very general solution of Eq. (18) will then be given by the series

$$\psi(t,x) = \sum_{n=0}^{\infty} c_n u_n(x) e^{(2\pi i/h)E_n t} \tag{74}$$

in which $u_n(x)$ is the n^{th} characteristic function of the space factor equation and E_n the corresponding value of E. The c's are arbitrary constants which must be so chosen to insure the convergence of the series. Now Eq. (18) shows that $\psi(t, x)$ is in the most general case fully determined by its value at $t=0$. Thus Eq. (74) does give the most general solution of the problem provided that it can be fitted to an arbitrary complex initial function $f(x)$ which meets the boundary conditions. Hence it must be proved that such a function can always be expanded into the infinite series

$$f(x) = \sum_{n=0}^{\infty} c_n u_n(x). \tag{75}$$

Such an expansion is most easily carried out in the case of a *normalized*

orthogonal system of functions $u_n(x)$.[34] Here we use the word *normalized* in the sense of Eq. (39) to indicate that for every value of n

$$\int_a^b u_n(x)u_n^*(x)dx=1\,, \tag{76}$$

where $x=a$, b are the boundaries of the region of definition of the functions. This equation can always be satisfied by suitable adjustment of the arbitrary multiplicative constant α. The term *orthogonal*[35] is applied to the series of functions $u_n(x)$ when

$$\int_a^b u_n(x)u_m^*(x)dx=0\,, \qquad n\neq m. \tag{77}$$

If the given wave functions form such a set, the coefficients in Eq. (75) may be determined like those in a Fourier series.† Multiplication of the equation by $u^*_\tau(x)$ and integration yields

$$c_\tau=\int_a^b f(x)u_\tau^*(x)dx\,; \qquad \tau=0,\,1,\,2,\cdots. \tag{78}$$

A slightly more general case is that in which the given set of functions is not orthogonal as it stands, but may be converted into such a set through multiplication by a suitable real common factor $[\bar{p}(x)]^{1/2}$. The condition that the modified functions form a normalized orthogonal system is then

$$\int_a^b \bar{p}(x)u_n(x)u_m^*(x)dx=\delta_{nm}\,, \tag{79}$$

where the symbol δ_{nm} stands for unity if $m=n$ and zero if $m\neq n$. Eq. (78) becomes

$$c_\tau=\int_a^b \bar{p}(x)f(x)u_\tau^*(x)dx\,. \tag{80}$$

The characteristic functions of the Sturm-Liouville problem are readily proved[36] to have this modified type of orthogonality, the function $\bar{p}$ being

[34] Cf. Courant-Hilbert, Chapter II; Riemann-Weber, Chapter VIII, 1, Chapter XII, 3.

[35] A continuous function of x may be regarded as a vector in a space of infinitely many dimensions and Eq. (77) is a limiting case of the condition for the orthogonality of two complex vectors U_n, U_m in a space of many dimensions. Cf. pp. 210, 211 below.

† Obviously if the expansion (75) is to be valid, the series of functions $u_n(x)$ must be *complete*, i.e., no function can exist not identically zero which is not in the series and is orthogonal to all members of the series. Cf. Courant-Hilbert, pp. 35–38.

[36] If λ_1 and λ_2 are two different characteristic values of Eq. (49), y_1, y_2 being the characteristic functions,

$$(\lambda_1-\lambda_2)\int_a^b \rho y_1y_2^*dx+\int_a^b \frac{d}{dx}(p[y_1'y_2^*-y_1y_2^{*\prime}])dx=0.$$

The second expression vanishes due to the boundary conditions and since $\lambda_1\neq\lambda_2$, the orthogonality condition follows at once. Cf. Courant-Hilbert, p. 239; Riemann-Weber, p. 366.

identical to a constant factor with the $\rho(x)$ of Eq. (49).‡ Moreover, the development of Eq. (75) is valid in every case [37] so that the series (74) does give the most general solution of the one-dimensional wave equation which meets the boundary conditions.

In the case of the linear oscillator equation (48) under consideration, ρ is a constant and we may normalize the wave functions in accordance with Eqs. (76) and (39) to preserve our original simple physical interpretation of $\psi\psi^* dx$ for the case where ψ has the monochromatic form $\psi = u_n e^{+(2\pi i/h)E_n t}$.

8. Continuous Spectrum

The above results have been generalized to include the case where there is a continuous as well as a discrete spectrum. In this case one of the boundaries is at infinity and a part of the series expansion (75) goes over into an integral analogous to the Fourier integral. Thus if $f(x)$ is a function which, together with its second derivative is continuous and quadratically integrable over the fundamental region $a < x < \infty$, it can be expanded in the absolutely and uniformly convergent representation[38]

$$f(x) = \sum_n c_n u_n(x) + \int_\epsilon^\infty c(E) u(x, E) dE. \tag{81}$$

Here $u(x, E)$ is the space factor for the energy E and ϵ is the minimum energy value for the continuous spectrum. (We assume that the constant of integration β (p. 179) is determined by the boundary condition at one end of the fundamental region so that, except for the multiplicative constant α, $u(x, E)$ depends on x and E only.)

The discrete functions $u_n(x)$ are normalized and their coefficients c_n determined as before. The normalization-orthogonality relation Eq. (79) does not apply to the characteristic functions of the continuous spectrum, however, since the integrals do not converge. This difficulty is met by the use of the modified condition that[39]

$$\int_a^\infty dx \bar{\rho}(x) u^*(x, \lambda) \int_{E_1}^{E_2} u(x, E) dE = F(\lambda) \quad \text{or} \quad 0 \tag{82}$$

according as $E_1 < \lambda < E_2$, or not. Then the formula for determining the coefficients $c(E)$ of the continuous spectrum is

$$c(E') = \frac{1}{F(E')} \frac{d}{dE'} \int_a^\infty dx \bar{\rho}(x) f(x) \int_\epsilon^{E'} u^*(x, \lambda) d\lambda. \tag{83}$$

For our present purpose $F(E')$ is arbitrary and Schrödinger sets it equal to

‡ ρ can always be eliminated by an appropriate change of the dependent variable as in Eq. 69.

[37] Courant-Hilbert, p. 278; Riemann-Weber VII §4, XIII, §1.

[38] H. Weyl, Math. Ann. **68,** 220 (1910).

[39] Cf. E. Fues, Ann. d. Physik (4) **81,** p. 281 (1926); E. Schrödinger, Ann. d. Physik (4) **81,** p. 109 (1926). Other forms of this condition are used by Oppenheimer, Zeits. f. Physik **41,** p. 268 (1927); Phys. Rev. **31,** p. 66 (1928).

unity. Oppenheimer, on the other hand, finds it convenient from the standpoint of statistical weights to give this function the constant value h.

The following heuristic procedure will serve to make the above formulas plausible. If we should try to determine the coefficients $c(E)$ by the same method as that used for the c_n's we would multiply Eq. (81) by $\bar{\rho}(x)u^*(x, E')$ and integrate formally over all values of x. It may happen, however, that the integral $\int_a^\infty \bar{\rho}(x)f(x)u^*(x, E')dx$ does not exist. Hence we must replace $u^*(x, E')$ by the "wave packet" $\int_\epsilon^{E'} u^*(x, \lambda)\, d\lambda$ which approaches zero rapidly for large values of x due to interference of its elements. The resulting equation is

$$\int_a^\infty dx\bar{\rho}(x)f(x)\int_\epsilon^{E'} u^*(x,\lambda)d\lambda = \sum_n c_n \int_a^\infty dx\bar{\rho}(x)u_n(x)\int_\epsilon^{E'} u^*(x,\lambda)d\lambda + \int_a^\infty \bar{\rho}(x)dx\int_\epsilon^{E'} u^*(x,\lambda)d\lambda\int_\epsilon^\infty c(E)u(x,E)dE. \tag{84}$$

Due to difficulty about interchanging the order of integrations in the right hand side of Eq. (84) we make use of the fact that the characteristic functions of the continuous spectrum may be regarded as limiting cases of discrete characteristic functions obtained by a suitable modification of the potential energy function. [E.g., if the potential energy $V_0(x)$ has the form indicated in Fig. 2 with a continuous spectrum, we may consider a modified problem in which

$$V = V_0(x) + V_1(x)$$

where V_1 vanishes if $x < x_1$ and has the value $V_1 = a(x - x_1)$ if $x > x_1$. Then the modified problem has no continuous spectrum since $V(\infty) = \infty$, but if x_1 is allowed to increase without limit the energy levels above the critical value $V_0(\infty)$ become indefinitely close together and the wave functions approximate more and more closely to those of the original problem.] On this basis we may prove (a) that the sum which forms the first term of the right hand member is zero, and (b) that $\int_a^\infty dx\bar{\rho}u^*(x, \lambda)\int_{E_2}^{E_1} c(E)u(x, E)dE$ vanishes unless $E_1 < \lambda < E_2$ and is independent of $E_1 - E_2$. The value of the integral is then proportional to $c(\lambda)$ and by the introduction of a suitable normalization factor can be reduced to $c(\lambda)F(\lambda)$ where F is arbitrary [Cf. Eq. (82)]. Then Eq. (84) yields

$$\int_a^\infty dx\bar{\rho}f(x)\int_\epsilon^{E'} u^*(x,\lambda)d\lambda = \int_\epsilon^{E'} c(\lambda)F(\lambda)d\lambda.$$

Differentiating with respect to E' we obtain Eq. (83).

An important auxiliary relation similar to Eq. (45) is

$$\int_a^\infty ff^*dx = \sum_n c_n c_n^* + \int_\epsilon^\infty c(E)c^*(E)dE. \tag{85}$$

From the above development it follows that the most general solution of the one-dimensional wave equation for $\psi(x, t)$, viz.,

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2\mu}{h^2}\left[\frac{h}{2\pi i}\frac{\partial}{\partial t} - V\right]\psi = 0, \tag{86}$$

in harmony with the boundary condition is

$$\psi(x,t) = \sum_n c_n u_n e^{(2\pi i/h)E_n t} + \int_\epsilon^\infty c(E)u(x,E)e^{(2\pi i/h)Et}dE. \tag{87}$$

9. Orthogonality and Expansion Properties of the Characteristic Functions of the Two-Body Problem

By means of Green's theorem one may prove that any two characteristic functions of the Schrödinger equation (60) corresponding to different energy levels are orthogonal to one another and to any packet formed from characteristic functions belonging to the continuous spectrum.[40] Furthermore, if the variables are separated, mutual orthogonality relations are readily established which, when combined with a suitable normalization, are expressed by the formulas

$$\int_0^{2\pi} \Phi_m \Phi_{m'}^* d\phi = \delta_{m,m'} : \quad \int_0^{\pi} \Theta_{l,m} \Theta_{l',m} \sin\theta d\theta = \delta_{l,l'}$$
$$\int_0^{\infty} R_{n,l} R_{n,l} r^2 dr = \int_0^{\infty} \Re_{n,l} \Re_{n',l} dr = \delta_{n,n'}. \tag{88}$$

Hence

$$\int\int\int_{\infty} u_{n,l,m} u_{n',l',m'}^* dx dy dz = \int\int\int_{\infty} R_{n,l} \Theta_{l,m} \Phi_m R_{n',l'} \Theta_{l',m'} \Phi_{m'}^* r^2 \sin\theta dr d\theta d\phi$$
$$= \delta_{n,n'} \delta_{l,l'} \delta_{m,m'}. \tag{89}$$

It can also be proved that an arbitrary function of θ and ϕ which, together with its first and second derivatives, is continuous over the entire sphere may be developed into an absolutely and uniformly convergent series of spherical harmonics[41] $\Theta_{l,m}\Phi_m$. Combining this fact with the corresponding theorem for $R(r)$ which follows from the Sturm-Liouville character of the differential equation (62), we conclude that a suitably continuous arbitrary function of x, y, z or of r, θ, ϕ may be expanded in the form

$$f(x,y,z) = \sum_{n=l+1}^{\infty} \sum_{l=|m|}^{\infty} \sum_{m=-\infty}^{+\infty} c_{n,l,m} R_{n,l} \Theta_{l,m} \Phi_m$$
$$+ \int_0^{\infty} dE \left\{ \sum_{l=|m|}^{\infty} \sum_{m=-\infty}^{+\infty} c_{l,m}(E) R_l(E,r) \Theta_{l,m} \Phi_m \right\}. \tag{90}$$

As in the simpler case of the one-dimensional oscillator, we have only to insert appropriate time factors in the right hand member of Eq. (90) to obtain the most general solution of the wave equation for the internal motion of the two-body problem in the form (18).

Similar orthogonality relations and expansion properties hold, no doubt, for most, if not all, of the characteristic value problems of the wave mechan-

[40] Cf. Courant-Hilbert, p. 255. Two functions u_n, u_m of the space coordinates x, y, z are by definition orthogonal in the region consisting of all space if

$$\iiint_{\infty} u_n u_m^* dx dy dz = 0, \qquad n \neq m.$$

[41] Cf. Courant-Hilbert, p. 422.

ics, though the writer knows of no comprehensive and rigorous discussion of the subject. The plausibility of these relations may be made evident in various ways which lie outside the scope of this review and is so great that they have been assumed without question by most theoretical physicists. They are of the greatest importance not only for expressing the most general solution of the basic wave equation (38) in terms of characteristic functions defined by the space factor equation (37), but also in the development of perturbation theory and in the study of the problem of the emission of radiation by matter.

SECTION 3. OPERATORS AND MOMENTA IN THE WAVE MECHANICS

1. Mean Values of Functions of the Coordinates

Mean values play an important rôle in any statistical theory. In fact we may say that the function of any statistical theory is to determine mean values of one kind and another. For example, $\psi\psi^* d\tau$ may be regarded as a mean value for the number of particles in the volume element $d\tau$.

To find the mean value of any function of the coordinates $f(q)$ we have to multiply the probability that the system lies in any particular element in coordinate space by the corresponding value of $f(q)$ and sum. Thus

$$\overline{f(q)} = \int f(q)\psi\psi^* d\tau. \tag{91}$$

As an example of the use of this type of mean value we may cite the rigorous derivation of Newton's laws of motion for wave packets in the form

$$\mu\ddot{\bar{x}} = -\frac{\overline{\partial V}}{\partial x} \tag{92}$$

by Ehrenfest.[41a]

2. The Linear Momentum Operator

Consider next the average values of the components of momentum. In the case of a single particle we have already shown that by analysing $\psi(t, x, y, z)$ into plane waves we may derive a wave function $Q(t, p_x, p_y, p_z)$ such that $QQ^* dp_x dp_y dp_z$ gives the probability that the momentum vector terminates in the element of volume $dp_x dp_y dp_z$ in momentum space. Then clearly the mean value of any component of momentum, say p_x is

$$\overline{p_x} = \int\int\int_{\infty} p_x QQ^* dp_x dp_y dp_z. \tag{93}$$

By means of the operator $(-h/2\pi i)\partial/\partial x$ we can replace the above mode of averaging by a more direct one which does not involve the evaluation of the Q function. When applied to a wave function corresponding to a single

[41a] P. Ehrenfest, Zeits. f. Physik **45**, p. 455 (1927). Ehrenfest's proof for the one-dimensional case has been extended to the general case of a conservative system composed of n particles by Ruark, Phys. Rev. **31**, p. 533 (1928). Cf. also Sommerfeld "Ergänzungsband," pp. 287–9.

definite value of p_x [and hence of the form $f(t, y, z)e^{-(2\pi i/h)xp_x}$] the operator yields the relation

$$-\frac{h}{2\pi i}\frac{\partial\psi}{\partial x}=p_x\psi. \tag{94}$$

Hence, differentiating Eq. (42) with respect to x, we have

$$-\frac{h}{2\pi i}\frac{\partial\psi}{\partial x}=\frac{1}{h^{3/2}}\int\int\int_{\infty} p_xQe^{(2\pi i/h)(xp_x+yp_y+zp_z)}dp_xdp_ydp_z \tag{96}$$

Now by a slight extension of the theorem of Eq. (45)[42] we may show that if $\psi_1, Q_1; \psi_2, Q_2$ are two pairs of functions which satisfy the relations of Eqs. (42)and (43)

$$\int\int\int_{\infty}\psi_1\psi_2^*dxdydz=\int\int\int_{\infty}Q_1Q_2^*dp_xdp_ydp_z. \tag{97}$$

Hence, identifying ψ_1 with $-(h/2\pi i)\partial\psi/\partial x$ and ψ_2 with ψ, we obtain

$$\int\int\int_{\infty}\psi^*\left(-\frac{h}{2\pi i}\frac{\partial}{\partial x}\right)\psi dxdydz=\int\int\int_{\infty}Q^*p_xQdp_xdp_ydp_z=\bar{p}_x. \tag{98}$$

Thus the mean value of each of the components of momentum may be evaluated by a rule formally the same as that used for $f(q)$ but with the substitution of the *operator* $-(h/2\pi i)\partial/\partial q$ for the momentum p conjugate to the coordinate q. Similarly we may determine the mean value of p^n by the rule

$$\overline{p^n}=\int\int\int_{\infty}\psi^*\left(-\frac{h}{2\pi i}\frac{\partial}{dq}\right)^n\psi dxdydz. \tag{99}$$

The above mean value formulas are valid for a particle moving in a force field as well as for a free particle and are readily extended to cover a system of particles. In the latter case the Fourier analysis may be applied to one coordinate only as follows.[43]

$$\left.\begin{aligned}\psi(t,x_1,x_2,\cdots,x_{3n})&=h^{-1/2}\int_{-\infty}^{+\infty}\phi(t,x_1,\cdots,p_k,\cdots,x_{3n})e^{-(2\pi i/h)x_kp_k}dp_k\\ \phi(t,x_1,\cdots,p_k,\cdots,x_{3n})&=h^{-1/2}\int_{-\infty}^{+\infty}\psi(t,x_1,\cdots,x_k,\cdots,x_{3n})e^{(2\pi i/h)x_kp_k}dx_k\end{aligned}\right\} \tag{100}$$

$$\int_{\infty}\psi^*\left(-\frac{h}{2\pi i}\frac{\partial}{\partial x_k}\right)\psi dx_1\cdots dx_{3n}$$

$$=\int_{\infty}\phi^*p_k\phi dx_1\cdots dp_k\cdots dx_{3n}=\overline{p_k}. \tag{101}$$

[42] The mode of procedure is suggested in Riemann-Weber, p. 299 in connection with the proof of Eq. (16).

[43] It follows from the analysis of Sect. 1, that the fundamental formula (12) can be used to correlate the components of the momentum of the individual particles in such a system with the corresponding components of the wave-number in configuration space.

Evidently the total momentum in the direction of the x axis may be averaged by similar use of the operator $\sum_{l=1}^{n} (-h/2\pi i)\partial/\partial x_l$.

As Sommerfeld[44] has pointed out, the use of the operator $(-h/2\pi i)\partial/\partial q_k$ for the momentum p_k conjugate the coordinate q_k is intimately related to a fundamental theorem derived by Schrödinger[45] and interpreted by him as a statement of the law of the conservation of electricity. In its simplest form for a single particle in three dimensions using Cartesian coordinates this theorem is

$$\frac{\partial}{\partial t}(\psi\psi^*) = \frac{h}{4\pi\mu i} \operatorname{div} [\psi \operatorname{grad} \psi^* - \psi^* \operatorname{grad} \psi]. \tag{102}$$

Multiplying through by the charge e and interpreting $e\psi\psi^*$ as the charge density (statistical mean), Eq. (102) takes the form of the equation of continuity with the vector current density $\boldsymbol{S}$ defined by

$$\boldsymbol{S} = \frac{eh}{4\pi\mu i}[\psi \operatorname{grad} \psi^* - \psi^* \operatorname{grad} \psi]. \tag{103}$$

Multiplication of $\boldsymbol{S}$ by μ/e should give the vector momentum per unit volume, and integration over all coordinate space should give the total average momentum. In this way we find

$$\overline{p_x} = \frac{h}{4\pi i} \int_{\infty} \int \int \left[\psi \frac{\partial \psi^*}{\partial x} - \psi^* \frac{\partial \psi}{\partial x}\right] dx dy dz. \tag{104}$$

The equivalence of Eq. (104) and our previous expression for $\overline{p_x}$ [Eq. (98)] may be proved by partial integration of Eq. (104) or by reference to Eqs. (94) and (95).

The theorem of Eq. (102) is given by Schrödinger in a general form applicable to a system of n particles described by means of any sort of generalized coordinates. Sommerfeld has used it to resolve Eq. (92) into the pair of equations

$$\overline{p_x} = \mu\dot{\bar{x}} \quad (105) ; \qquad \dot{\overline{p_x}} = -\overline{\partial V/\partial x}. \tag{106}$$

3. The Angular Momentum Operator

The angular momentum of a system of particles is defined in the classical mechanics by the vector formula

$$\boldsymbol{M} = \sum_{k=1}^{n} \boldsymbol{r}_k \times \boldsymbol{p}_k. \tag{107}$$

As this formula when expressed in terms of a Cartesian coordinate system involves all the coordinates and all the momenta it might appear at first glance that the above methods of computing the mean value of $\boldsymbol{M}$ would

[44] Sommerfeld "Ergänzungsband," pp. 284–5.

[45] E. Schrödinger, Ann. d. Physik (4) **81**, p. 136 (1926).

break down due to the impossibility of assigning exact simultaneous values to the coordinates and conjugate momenta. Ignoring this possible difficulty for the moment, however, we observe that by means of the Cartesian expressions for the components of $\boldsymbol{M}$ in terms of the components of the linear momentum we may write out operators to be associated with M_x, M_y, M_z, respectively. Thus, in the case of a single particle

$$M_z = xp_y - yp_x \rightarrow -\frac{h}{2\pi i}\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right). \tag{108}$$

Now the definition of $\boldsymbol{M}$ given in Eq. (107) is, of course restricted to wave packets in the limiting case where the uncertainties in coordinates and momenta can be ignored. Outside this special realm of the classical mechanics we are at liberty to define the components of angular momentum as we please, provided only that our definitions are in agreement with Eq. (107) in cases where the latter is valid. Hence, in harmony with Eq. (94) we define M_z for ψ functions which give it a unique numerical value by means of the equation

$$-\frac{h}{2\pi i}\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right)\psi = \alpha\psi. \tag{109}$$

If the quantity α is independent of x, y, z, we indentify it with M_z. Otherwise we assign no definite value of M_z to the wave function under consideration but assume that it has a certain probability of taking on any one of a variety of values like the components of linear momentum for a wave packet. As a means for determining the probability amplitude for M_z in such cases we adopt the scheme of analyzing the wave function into a linear combination of functions which satisfy Eq. (109) and the usual boundary conditions for some value of α independent of the space coordinates. Thus Eq. (109) is to be made the basis of a characteristic value problem like the Schrödinger equation (15). The characteristic values of α are the possible values of M_z. The characteristic functions are the types of wave function which have definite angular momenta along the z axis.

Before formulating a mean value theorem for M_z we must give the differential equation (109) a brief examination. If we introduce the same spherical coordinates as in the two body problem [Sect. 1, 6] the operator $(h/2\pi i)[x(\partial/\partial y) - y(\partial/\partial x)]$ becomes $(h/2\pi i)\partial/\partial\phi$ and our differential equation takes the simple form

$$\frac{h}{2\pi i}\frac{\partial\psi}{\partial\phi} + M_z\psi = 0. \tag{110}$$

Its solution is

$$\psi = \chi(t, r, \theta)e^{-(2\pi i/h)M_z\phi}. \tag{111}$$

Continuity of ψ regarded as a function of x, y, and z requires that $2\pi M_z/h$ be restricted to integral characteristic values. M_z has no continuous spectrum

This is the origin of the quantization of angular momentum as it occurs in the Bohr theory. From our present point of view non-integral values of $2\pi M_z/h$ have no meaning.

As an arbitrary continuous function of x, y, z expressed in terms of the spherical coordinates r, θ, ϕ has the period 2π in ϕ, it follows that such a function can be expanded into the series of characteristic functions

$$\psi(t,r,\theta,\phi)=\sum_{m=-\infty}^{+\infty}\chi_m(t,r,\theta)e^{-im\phi}. \tag{112}$$

$\chi_m(t,r,\theta)$ plays the rôle of probability amplitude for the three independent variables r, θ, M_z. To get the mean value of M_z for the given ψ function we use the same procedure as for $\bar{p}_z$.

$$\overline{M_z}=\sum_m\int\chi^*{}_m M_z\chi_m d\tau=\int\psi^*\left(-\frac{h}{2\pi i}\frac{\partial}{\partial\phi}\right)\psi d\tau \tag{113}$$

where $d\tau$ is the element of volume $dxdydz$ or $r^2\sin\theta\, drd\theta d\phi$.

The definition of M_z involved in equation (110) can now be justified by applying it to a wave packet having a precisely defined position and momentum as required by the classical mechanics. In the case of such a packet one can readily prove that

$$\overline{xp_y}=\int\psi^*\left(-x\frac{h}{2\pi i}\frac{\partial}{\partial y}\right)\psi d\tau\cong\bar{x}\bar{p}_y.$$

Hence $\overline{M_z}$ is approximately equal to $\bar{x}\bar{p}_y-\bar{y}\bar{p}_x$. As the classical values of coordinates and momenta are identical with the mean values for the packet, this shows that our definition reduces to the classical one in the realm of validity of the older theory.

Sommerfeld[46] has proved the additional theorem that

$$\dot{\boldsymbol{M}}=\boldsymbol{r}\times\text{grad } V=\textit{mean applied torque.} \tag{114}$$

This he calls the integral form of the angular momentum law for the wave mechanics.

The possibility of assigning an exact value to M_z and hence also to M_x or to M_y, does not carry with it the possibility of assigning an exact value to the vector angular momentum $\boldsymbol{M}$, for simultaneous solutions of Eq. (110) and the corresponding equations for M_x and M_y do not exist. On the other hand, unique values for $\boldsymbol{M}^2=M_x{}^2+M_y{}^2+M_z{}^2$ are possible. The corresponding operator is

$$-\frac{h^2}{4\pi^2}\left[\left(y\frac{\partial}{\partial z}-z\frac{\partial}{\partial y}\right)^2+\left(z\frac{\partial}{\partial x}-x\frac{\partial}{\partial z}\right)^2+\left(x\frac{\partial}{\partial y}-y\frac{\partial}{\partial x}\right)^2\right].$$

Introducing spherical coordinates, applying the above operator to ψ, and identifying the resulting expression with $-\boldsymbol{M}^2\psi$, we obtain the differential equation

[46] Sommerfeld "Ergänzungsband," p. 290.

$$\frac{h^2}{4\pi^2}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}\right]-\boldsymbol{M}^2\psi=0. \tag{115}$$

This is identical with the differential equation (63) whose solutions are the tesseral spherical harmonics. Thus one of the differential equations obtained on separating the variables in the two-body problem is identical with the condition that the wave functions for the various energy levels in that problem shall have definite values of $\boldsymbol{M}^2$ or of $|\boldsymbol{M}|$.[47] The characteristic values of $\boldsymbol{M}^2$ for the one-body problem or for the two-body problem are given by the formula

$$\boldsymbol{M}^2=\frac{h^2}{4\pi^2}l(l+1) \qquad l=0,1,2,\cdots \tag{116}$$

first derived by the matrix mechanics.[48] The extension of this formula to the problem of n particles presents no difficulty.

As Eqs. (110) and (111) are also valid for the special solutions of the two-body characteristic value problem of Sect. 2, 6 in which the variables are separated, we see that these solutions have unique values of M_z as well as of $\boldsymbol{M}^2$. As previously noted, M_x and M_y cannot have unique values for functions which make M_z unique. Their mean values are zero for the special solutions under consideration and the sum of the mean squares is

$$\overline{M_x{}^2}+\overline{M_y{}^2}=\overline{\boldsymbol{M}^2}-\overline{M_z{}^2}=\frac{h^2}{4\pi^2}[l(l+1)-m^2]. \tag{117}$$

4. The Energy Operators

The entire theory developed up to this point is based on the hypothesis that all ψ functions are linear combinations of functions of the form $u(t, q)e^{(2\pi i/h)Et}$. Hence if any particular wave function is monochromatic (i.e., has a unique energy value)

$$\frac{h}{2\pi i}\frac{\partial\psi}{\partial t}=E\psi. \tag{118}$$

[47] It is hardly necessary to note that if we start from the definition

$$M_x=\sum_{k=1,2}[y_k(p_z)_k-z_k(p_y)_k]$$

and introduce the coordinates of the center of gravity ξ, η, ζ together with the relative coordinates x, y, z as in Sect 2, 6, the operator for M_x splits into two parts, one of which is

$$-\frac{h}{2\pi i}\left(\eta\frac{\partial}{\partial\zeta}-\zeta\frac{\partial}{\partial\eta}\right)$$

and gives the angular momentum of the center of gravity, while the other is formally identical with the operator used in Eq. (110) for a single particle and gives the angular momentum of the two bodies about their center of gravity. Applying the same procedure to the other components of M we justify the direct application to the two-body problem of the operators developed above for a single particle.

[48] Born, Heisenberg, and Jordan, Zeits. f. Physik **35**, p. 557 (1925).

The mean value of E for any normalized wave function is readily shown to be

$$\overline{E} = \int_{\infty} \psi^* \left(\frac{h}{2\pi i} \frac{\partial}{\partial t} \right) \psi d\tau. \tag{119}$$

Thus the operator $(h/2\pi i)\partial/\partial t$ bears the same relation to the numerical values of E as $-(h/2\pi i)\partial/\partial x$ bears to the numerical values of p_x. [Classically E and $-t$ are canonically conjugate variables like p_x and x].

There is another operator, however, associated with the energy and derivable from the classical Hamiltonian function. In the case of a set of n particles moving in a conservative force field the expression for the energy in terms of coordinates and momenta (classical Hamiltonian) is

$$H = \sum_{k=1}^{3n} \frac{{p_k}^2}{2\mu_k} + V(x_1, \cdots, x_{3n}).$$

If we replace the p_k's by the corresponding operators we obtain a new operator which we will designate as $H(x_k, -\partial/\partial x_k)$. Thus[49]

$$H\left(x_k, -\frac{\partial}{\partial x_k} \right) = -\sum_{k=1}^{3n} \frac{h^2}{8\pi^2 \mu_k} \frac{\partial^2}{\partial {x_k}^2} + V. \tag{120}$$

If this operator is applied to ψ and the resulting expression is equated to $-(h/2\pi i)\partial\psi/\partial t$ one obtains

$$\left[\sum_{k=1}^{3n} \frac{1}{\mu_k} \frac{\partial^2}{\partial {x_k}^2} - \frac{8\pi^2 V}{h^2} + \frac{4\pi}{hi} \frac{\partial}{\partial t} \right] \psi = 0. \tag{121}$$

This is the basic equation of the Schrödinger theory (36). Eq. (36) is therefore equivalent to a statement of the identity of the two operators for E when applied to physically permissible wave functions. In view of this relation between the fundamental wave equation and the Hamiltonian function it is customary to write the differential equation in the symbolic form

$$H\psi = \frac{h}{2\pi i} \frac{\partial \psi}{\partial t}. \tag{122}$$

The disclosure of the intimate relation between the Hamiltonian function of classical theory and the operator H of Eq. (122) when Cartesian coordinates are used immediately raises the question of the relation between these expressions in other coordinates. This problem has been investigated by Schrödinger[50] and by Podolsky[51] who find that given the Hamiltonian func-

[49] V as an operator is interpreted to mean "multiply by the numerical value of the function V."

[50] E. Schrödinger, Ann. d. Physik (4) **79**, pp. 747, 748 (1926).

[51] B. Podolsky, Phys. Rev. **32**, p. 812 (1928). Cf. also P. A. M. Dirac, Proc. Roy. Soc. **A113**, p. 621 (1927). Schrödinger's method gives what Podolsky calls ψ_x, i.e. the ψ function of Eqs. (18) and (36) with the physical interpretation "$\psi_x \psi_x^* dx_1 \cdots dx_{3n}$ = probability of configuration in the element $dx_1 \cdots dx_{3n}$ of Cartesian coordinate space." Dirac's transformation theory introduces a different function, ψ_q, for the generalized coordinates $q_1, \cdots, q_{3n}$ such

tion in any system of coordinates one can deduce the corresponding operator H without the intermediate step of reverting to a Cartesian system. The operator H is frequently written as $H(q, \partial/\partial q)$ suggesting that it can be formed in any set of cordinates by the mere substitution of $-(h/2\pi i)\partial/\partial q$ for the component of momentum conjugate to q in the classical Hamiltonian. The problem is not quite so direct and simple as that, however, since the suggested procedure is ambiguous. Consider, for example, the kinetic energy term $p_r^2/2\mu$ of the classical Hamiltonian in spherical coordinates which goes over into the operator

$$\frac{-h^2}{8\pi^2\mu r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)$$

So long as p_r is a number the factor p_r^2 can be replaced by $[1/f(r)]p_r f(r)p_r$ with $f(r)$ arbitrary, but when we convert p_r into an operator we must use the definite form $r^{-2}p_r r^2 p_r$. Hence a rule is needed for choosing that particular way of writing the classical Hamiltonian which will yield the correct operator when one makes the substitution $p_k \rightarrow (-h/2\pi i)\partial/\partial q_k$.

The converse step of deducing the classical Hamiltonian function $H(q, p)$ from the wave equation in generalized coordinates involves no such difficulty.

Incidentally the possibility of deducing the wave equation in any set of coordinates from $H(q, p)$ and the above operator substitution suggests that we define numerical values of the generalized component of momentum conjugate to any coordinate q as we defined the numerical values of M_z in Eq. (110), i.e. that we ascribe the value α to p_k when ψ satisfies the differential equation

$$-\frac{h}{2\pi i}\frac{\partial}{\partial q_k}\psi(t, q_1, \cdots, q_{3n}) = \alpha\psi(t, q_1, \cdots, q_{3n}) \tag{123}$$

and fix the probability of various values of p_k for a normalized ψ function by an analysis of the function into a linear combination of the solutions

$$\psi = \chi(t, q_1, \cdots, q_{k-1}, q_{k+1}, \cdots, q_{3n})e^{-(2\pi i/h)q_k p_k}.$$

5. Conservation of Energy and Momentum.

In the classical mechanics the law of the conservation of energy ascribes to every conservative system a function of its coordinates and momenta which is called its energy and which is constant in time as the motion progresses. In the wave mechanics we have in general no single definite energy, but in general a distribution function giving the probability of various possible energy values. By "conservation of energy" we may therefore imply the constancy of this distribution function in time, or, what comes to the same thing, the constancy of the mean values of E and its various powers.

---

that $\psi_q\psi_q^* dq_1 \cdots dq_{3n}$ is the probability that the configuration lies in the element $dq_1 \cdots dq_{3n}$ of the corresponding coordinate space and Podolsky gives a method of writing out the Hamiltonian function such that the substitution of $-(h/2\pi i)\partial/\partial q_k$ for p_k yields directly the wave equation for ψ_q. ψ_q is equal to ψ_x multiplied by the square root of the Jacobean of the transformation from the Cartesian coordinate system to the system $q_1, \cdots, q_{3n}$.

In deriving this law it will be sufficient to deal with the special case of the linear oscillator. Eq. (87) shows that the most general physically admissible solution of the wave equation is a linear combination of monochromatic solutions. The relative probabilities of the different discrete energy values and the different elementary intervals of the continuous spectrum are given by the squares of the absolute values of the coefficients c_n and $c(E)$ respectively [Cf. Eq. (85)]. As these coefficients are constants the theorem is proved.

Defining the laws of the conservation of linear and angular momentum in the same way, we note that a sufficient condition for the conservation of any mechanical quantity α assocated with an operator O_α is that ψ shall be expansible into a linear combination of orthogonal functions each of which is monochromatic and also a characteristic solution of the equation

$$O_\alpha \psi = \alpha \psi . \tag{124}$$

This means that every non-degenerate solution of the wave equation (37) must also be a solution of Eq. (124) In the case of an energy level exhibiting r-fold degeneracy there exist an infinite number of sets of r mutually orthogonal solutions of Eq. (37) for the given energy. Each of these sets is derivable from any other by a suitable change of axes or by taking suitable linear combinations of the other solutions. In this case the conservation law holds for α if there exists some one set of r orthogonal solutions of Eq. (37) each member of which is also a solution of Eq. (124). In the case of a free particle the plane waves of Eq. (40) have the required characteristics for the conservation of linear momentum. In the case of the two-body problem referred to its center of gravity as origin, the solutions obtained in Sect. 2, 6 by separation of the variables in spherical coordinates have the required characteristics to prove the conservation of the z component of angular momentum Evidently a rotation of axes will suffice to give new solutions exhibiting the conservation of M_x and M_y.

Sommerfeld[52] defines what he calls the differential form of the "surface law" or law of the conservation of angular momentum by the equation

$$(H-E)\boldsymbol{M}\psi = 0. \tag{125}$$

in which $\boldsymbol{M}$ is the operator $iM_x+jM_y+kM_z$. This equation is derivable from Eq. (37) if the operators H and $\boldsymbol{M}$ commute, i.e., if $\boldsymbol{M}H\psi = H\boldsymbol{M}\psi$. It shows that if ψ is a solution of Eq. (37), $\boldsymbol{M}\psi$ is also a solution. Then if Eq. (37) has the r mutually orthogonal solutions $\psi_1, \psi_2, \cdots \psi_r$

$$\boldsymbol{M}\psi_n = \sum_{k=1}^{r} c_{nk}\psi_k . \tag{126}$$

Sommerfeld's form of the conservation law reduces to that given above if the operators M_x, M_y, M_z, or $\boldsymbol{M}^2$ are substituted for the vector operator $\boldsymbol{M}$. In each of these cases the required set of simultaneous solutions of Eq.

[52] Sommerfeld "Ergänzungsband," p. 295.

(37) and Eq. (124) are derivable from the relation (126). In the case of the vector operator $\boldsymbol{M}$, however, the equation (124) has no solutions except the trivial one where α is zero. Hence the above statement of the conservation law in terms of the distribution function applies to $\boldsymbol{M}$ only in so far as it is true for each of the components of $\boldsymbol{M}$ taken separately.

The constancy in time of the mean value of the quantity α may be proved if the operator O_α commutes with $H(q_k, -\partial/\partial q_k)$.

SECTION 4.* PERTURBATION THEORY

1. The Characteristic Value Problem and Perturbations which do not Involve the Time.

In the quantum mechanics as in the Bohr theory, perturbation methods are of fundamental importance due to the fact that so few problems can be rigorously solved by direct attack. The fundamental idea of these methods is that by starting from an approximate solution of the problem in hand one may compute by "hammer and tongs" a series of corrections designed to improve the degree of approximation. The procedure is easy or difficult according to one's success in choosing a happy starting point. The successive approximations may not converge on an exact solution of the problem in hand, but usually the first few steps do yield an appreciable improvement on the initial wave functions.

Consider first the solution of a typical Schrödinger characteristic value problem by the perturbation method. The best type of initial approximation consists in a complete[53] orthogonal function system in which one may expand any solution of the actual problem. Such a system may be obtained by the rigorous solution of a simplified related characteristic value problem[54] whose equation we write in the symbolic form

$$(H_0 - E)u = 0. \tag{127}$$

We call this the unperturbed equation and designate the corresponding unperturbed characteristic functions and characteristic values by $u_k{}^0$ and $E_k{}^0$ respectively. Let the differential equation to be solved (the perturbed differential equation) be of the form

$$(H - E)u \equiv (H_0 + H_1 - E)u = 0. \tag{128}$$

A formal solution of this equation can be obtained by the expansion of u in terms of the $u_k{}^0$'s as follows:

$$u = \sum_n c_n u_n{}^0 + \int_\epsilon^\infty c(E) u_E{}^0 dE. \tag{129}$$

Here $u_E{}^0$ is a wave function associated with the continuous spectrum of Eq. (127). Due to degeneracy in the wave functions of the continuous spectrum the integrand of the above integral will usually be a sum as in Eq. (90) but to avoid undue complication we use the above simplified

* Mostly machinery—a necessary evil!

[53] Cf. Note †, p. 187.

[54] Cf. Section 2.

form of expansion. In fact we shall omit the integral entirely in carrying through the detailed application of Eq. (129) to the solution of Eq. (128). This is permissible in view of the fact that in some cases Eqs. (127) and (128) have no continuous spectrum while in other cases a good approximation to the lower characteristic forms of u can be obtained without considering the continuous spectrum.

Let the function Hu_n^0 have the corresponding expansion

$$Hu_n^0 = \sum_m H(n,m)u_m^0 + \int_\epsilon^\infty H(n,E)\, u_E^0 dE. \tag{130}$$

The two-dimensional array of coefficients $H(n, m)$ forms a matrix which we designate by $\{H(n, m)\}$. The elements of this matrix are given by the formula

$$H(n,m) = \int_\infty u_m^{0*} H u_n^0 d\tau. \tag{131}$$

$\{H(n, m)\}$ has what is called the Hermitian character, i.e., the symmetrical elements $H(m, n)$ and $H(n, m)$ are conjugate complex quantities. The proof of this statement will be given in §3 of this section. Since the operator $H(q, -\partial/\partial q$ is real, it follows that

$$H(m,n) = H^*(n,m) = \int_\infty u_m^0 H u_n^{0*} d\tau. \tag{132}$$

A matrix is a function of two independent variables (sometimes two sets of independent variables) which take on only discrete values. The coefficients $H(n,E)$ appearing in Eq. (130) and the additional coefficients $H(E',m)$, $H(E',E)$ obtained from the expansion of $Hu_{E'}^0$ may be combined with the $H(n,m)$'s to form a single function of two sets of independent variables, which take on discrete values in certain regions and continuous ones in others. Each set must of course be just sufficient to pick out one member of the complete orthogonal set of wave functions under consideration. Consider, for example, the important case where H_0 is the Hamiltonian operator for the central force field problem of Eq. (60). The arbitrary ordinal number n used above can then be replaced by the three variables M_z, M^2, E, or, using the quantum numbers for the first two variables, by m, l, E. The function $H(n,m)$ then merges with $H(n,E)$, $H(E',m)$ and $H(E',E)$ above to form a single function $H(m', l', E'; m'', l'', E'')$ in which m', l', m'', l'' take on only discrete values while E' and E'' take on either real or discrete values according to the region under consideration. Such functions form the stock in trade of Dirac's q-number theory. Due to the manner in which they are used and to their symmetrical Hermitian character expressed by the relation

$$H^*(m'', l'', E''; m', l', E') = H(m', l', E'; m'', l'', E''),$$

such functions may be regarded as generalized matrices.

Introducing the expansion (129) into the wave equation (128) and neglecting the continuous spectrum, we have

$$\sum_n c_n (Hu_n^0 - Eu_n^0) = 0.$$

Expanding Hu_n in turn by Eq. (130), and neglecting the continuous spectrum again we obtain

$$\sum_n c_n \left[\sum_m H(n,m)u_m^0 - Eu_n^0 \right] = 0.$$

definite value of p_x [and hence of the form $f(t, y, z)e^{-(2\pi i/h)xp_x}$] the operator yields the relation

$$-\frac{h}{2\pi i}\frac{\partial\psi}{\partial x}=p_x\psi. \tag{94}$$

Hence, differentiating Eq. (42) with respect to x, we have

$$-\frac{h}{2\pi i}\frac{\partial\psi}{\partial x}=\frac{1}{h^{3/2}}\int\int\int_{\infty} p_xQe^{(2\pi i/h)(xp_x+yp_y+zp_z)}dp_xdp_ydp_z \tag{96}$$

Now by a slight extension of the theorem of Eq. (45)[42] we may show that if $\psi_1, Q_1; \psi_2, Q_2$ are two pairs of functions which satisfy the relations of Eqs. (42)and (43)

$$\int\int\int_{\infty}\psi_1\psi_2^*dxdydz=\int\int\int_{\infty}Q_1Q_2^*dp_xdp_ydp_z. \tag{97}$$

Hence, identifying ψ_1 with $-(h/2\pi i)\partial\psi/\partial x$ and ψ_2 with ψ, we obtain

$$\int\int\int_{\infty}\psi^*\left(-\frac{h}{2\pi i}\frac{\partial}{\partial x}\right)\psi dxdydz=\int\int\int_{\infty}Q^*p_xQdp_xdp_ydp_z=\bar{p}_x. \tag{98}$$

Thus the mean value of each of the components of momentum may be evaluated by a rule formally the same as that used for $f(q)$ but with the substitution of the *operator* $-(h/2\pi i)\partial/\partial q$ for the momentum p conjugate to the coordinate q. Similarly we may determine the mean value of p^n by the rule

$$\overline{p^n}=\int\int\int_{\infty}\psi^*\left(-\frac{h}{2\pi i}\frac{\partial}{dq}\right)^n\psi dxdydz. \tag{99}$$

The above mean value formulas are valid for a particle moving in a force field as well as for a free particle and are readily extended to cover a system of particles. In the latter case the Fourier analysis may be applied to one coordinate only as follows.[43]

$$\left.\begin{aligned}\psi(t,x_1,x_2,\cdots,x_{3n})&=h^{-1/2}\int_{-\infty}^{+\infty}\phi(t,x_1,\cdots,p_k,\cdots,x_{3n})e^{-(2\pi i/h)x_kp_k}dp_k\\ \phi(t,x_1,\cdots,p_k,\cdots,x_{3n})&=h^{-1/2}\int_{-\infty}^{+\infty}\psi(t,x_1,\cdots,x_k,\cdots,x_{3n})e^{(2\pi i/h)x_kp_k}dx_k\end{aligned}\right\} \tag{100}$$

$$\int_{\infty}\psi^*\left(-\frac{h}{2\pi i}\frac{\partial}{\partial x_k}\right)\psi dx_1\cdots dx_{3n}$$

$$=\int_{\infty}\phi^*p_k\phi dx_1\cdots dp_k\cdots dx_{3n}=\overline{p_k}. \tag{101}$$

[42] The mode of procedure is suggested in Riemann-Weber, p. 299 in connection with the proof of Eq. (16).

[43] It follows from the analysis of Sect. 1, that the fundamental formula (12) can be used to correlate the components of the momentum of the individual particles in such a system with the corresponding components of the wave-number in configuration space.

Evidently the total momentum in the direction of the x axis may be averaged by similar use of the operator $\sum_{l=1}^{n} (-h/2\pi i)\partial/\partial x_l$.

As Sommerfeld[44] has pointed out, the use of the operator $(-h/2\pi i)\partial/\partial q_k$ for the momentum p_k conjugate the coordinate q_k is intimately related to a fundamental theorem derived by Schrödinger[45] and interpreted by him as a statement of the law of the conservation of electricity. In its simplest form for a single particle in three dimensions using Cartesian coordinates this theorem is

$$\frac{\partial}{\partial t}(\psi\psi^*) = \frac{h}{4\pi\mu i} \operatorname{div} [\psi \operatorname{grad} \psi^* - \psi^* \operatorname{grad} \psi]. \tag{102}$$

Multiplying through by the charge e and interpreting $e\psi\psi^*$ as the charge density (statistical mean), Eq. (102) takes the form of the equation of continuity with the vector current density $\boldsymbol{S}$ defined by

$$\boldsymbol{S} = \frac{eh}{4\pi\mu i}[\psi \operatorname{grad} \psi^* - \psi^* \operatorname{grad} \psi]. \tag{103}$$

Multiplication of $\boldsymbol{S}$ by μ/e should give the vector momentum per unit volume, and integration over all coordinate space should give the total average momentum. In this way we find

$$\overline{p_x} = \frac{h}{4\pi i} \int_{\infty} \int \int \left[\psi \frac{\partial \psi^*}{\partial x} - \psi^* \frac{\partial \psi}{\partial x}\right] dx dy dz. \tag{104}$$

The equivalence of Eq. (104) and our previous expression for $\overline{p_x}$ [Eq. (98)] may be proved by partial integration of Eq. (104) or by reference to Eqs. (94) and (95).

The theorem of Eq. (102) is given by Schrödinger in a general form applicable to a system of n particles described by means of any sort of generalized coordinates. Sommerfeld has used it to resolve Eq. (92) into the pair of equations

$$\overline{p_x} = \mu\dot{\bar{x}} \quad (105) ; \qquad \dot{\overline{p_x}} = -\overline{\partial V/\partial x}. \tag{106}$$

3. The Angular Momentum Operator

The angular momentum of a system of particles is defined in the classical mechanics by the vector formula

$$\boldsymbol{M} = \sum_{k=1}^{n} \boldsymbol{r}_k \times \boldsymbol{p}_k. \tag{107}$$

As this formula when expressed in terms of a Cartesian coordinate system involves all the coordinates and all the momenta it might appear at first glance that the above methods of computing the mean value of $\boldsymbol{M}$ would

[44] Sommerfeld "Ergänzungsband," pp. 284–5.

[45] E. Schrödinger, Ann. d. Physik (4) **81**, p. 136 (1926).

As the next step in the development of the perturbation theory we write the Hamiltonian of the perturbed equation in the form

$$H = H_0 + \lambda h_1 \tag{140}$$

where λ is a parameter which we arbitrarily introduce into the problem if necessary and which may eventually be given the value unity. The wave functions and energy values will then depend upon λ and by allowing this parameter to approach zero we can carry the perturbed problem over into the unperturbed problem. It is now plausible to assume that the energy values and the coefficients c_{kn} which determine the wave functions are expansible in power series in λ.

$$\begin{aligned} E_k &= E_k^{(0)} + \lambda E_k^{(1)} + \lambda^2 E_k^{(2)} + \cdots , \\ c_{kn} &= c_{kn}^{(0)} + \lambda c_{kn}^{(1)} + \lambda^2 c_{kn}^{(2)} + \cdots . \end{aligned} \tag{141}$$

These expansions need not be valid but we proceed on the hopeful assumption that they are permissible.[59]

The first step in the evaluation of the coefficients in the above expansions is to note that the $E_k^{(0)}$'s must be the energy levels of the unperturbed system. If the perturbed energy levels are numbered in the same way as the unperturbed, we may identify $E_k^{(0)}$ with E_k^0. Also the limiting value of the perturbed wave function in the case where becomes equal to zero, i.e., $\sum_n c_{kn}^{(0)} u_n^0$ must be a solution of the unperturbed equation for the characteristic value E_k^0. *If E_k^0 is non-degenerate*, as we shall assume for the present, the above expression has the unique value u_k^0 and we conclude that $c_{kn}^{(0)} = \delta_{kn}$.

Using the values thus obtained for $E_k^{(0)}$ and $c_{kn}^{(0)}$, inserting the series (141) into Eqs. (139), replacing $H_1(n,m)$ by $\lambda h_1(n,m)$, and equating the coefficients of the first and second powers of λ to zero, one obtains

$$\lambda E_k^{(1)} = \lambda h_1(k,k) = H_1(k,k) \ ; \tag{142}$$

$$c_{kk}^{(1)} = 0 \ ; \quad \lambda c_{kn}^{(1)} = \frac{H_1(k,n)}{E_k^0 - E_n^0}, \quad n \neq k \ ; \tag{143}$$

$$\lambda^2 E_k^{(2)} = \lambda^2 \sum_n c_{kn}^{(1)} h_1(n,k) = \sum_n \frac{H_1(k,n) H_1^*(k,n)}{E_k^0 - E_n^0} \ . \tag{144}$$

These equations carry the approximation far enough for most purposes.

[59] As an example of an instance in which the above expansions are not rigorously valid the case of the Stark effect as worked out by Schrödinger may be mentioned. If one assumes a uniform electric field extending to infinity in all directions the perturbing potential is infinite at infinite points in the direction of the negative electric force and no monochromatic solutions of the perturbed problem exist. Nevertheless the solutions obtained by Schrödinger using the above series do agree with the experimental observations Evidently the mathematical difficulty here is introduced by an excessive idealization of the experimental conditions for observing the Stark effect.

The form of perturbation theory developed in this article is due to Schrödinger, though he does not separate the two steps, expansion in terms of the unperturbed wave functions and development in terms of the parameter, as we do. Cf. E. Schrödinger, Ann. d. Physik (4) **80,** p. 437, (1926). The conditions for the convergence of the series (141) have been worked out by A. H. Wilson. Cf. Wilson's papers, Proc. Roy. Soc. **A122,** p. 589 and **A124,** p. 176 (1929).

Eq. (142) states that in first approximation the change in the energy due to the perturbing Hamiltonian operator H_1 is equal to the "mean value of H_1" averaged over the unperturbed function u_k^0. This is the wave mechanical equivalent of a familiar corresponding theorem of the classical mechanics.[60] Eqs. (143) and (144) may be interpreted as the description of an interaction between the unperturbed wave functions and the unperturbed energy levels produced by the operator H_1. This interaction is small for widely separated energy levels both on account of the energy difference $E_k^0 - E_n^0$ in the denominators and on account of the fact that the matrix elements $H_1(k,n)$ for widely separated wave functions are normally quite small. This result justifies our previous statement that the lower energy levels of the discrete spectrum will be very slightly affected by interaction with the continuous spectrum.

The second order energy correction $\lambda^2 E_k^{(2)}$ is wholly due to interaction between the primary level E_k^0 under consideration and the other energy levels. The contribution of any other level E_n^0 is positive if the latter lies lower than E_k^0 and vice versa. Moreover, the perturbing effect of E_n^0 on E_k^0 is exactly equal and opposite to that of E_k^0 on E_n^0. Hence one may say that the perturbative Hamiltonian introduces a set of repulsive forces between the different energy levels.

Consider next the case where *the energy level E_k^0 is degenerate.* Let the $r+1$ energy levels $E_k^0, E_{k+1}^0, \cdots E_{k+r}^0$ associated with the wave functions $u_k^0, \cdots u_{k+r}^0$ fall together. Any perturbed wave function u_k such that

$$\lim_{\lambda \to 0} E_k = E_k^0$$

must reduce to the form

$$u_k^{(0)} = \sum_{n=k}^{k+r} c_{kn}^{(0)} u_n^0$$

when λ becomes equal to zero. Due to the degeneracy, however, the above sum does not reduce to a single term. The values of the $c_{kn}^{(0)}$'s and of the first order correction to the energy may be deduced from Eqs. (139) if the coefficients of the first power terms in λ are set equal to zero. By this means one obtains the $r+1$ equations

$$\left.\begin{aligned}
&c_{k,k}^{(0)}[h_1(k,k) - E_k^{(1)}] + c_{k,k+1}^{(0)} h_1(k,k+1) + \cdots c_{k,k+r}^{(0)} h_1(k,k+r) = 0 \\
&c_{k,k}^{(0)} h_1(k+1,k) + c_{k,k+1}^{(0)}[h_1(k+1,k+1) - E_k^{(1)}] \\
&\qquad + \cdots c_{k,k+r}^{(0)} h_1(k+1,k+r) = 0 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&c_{k,k}^{(0)} h_1(k+r,k) + c_{k,k+1}^{(0)} h_1(k+r,k+1) \\
&\qquad + \cdots c_{k,k+r}^{(0)}[h_1(k+r,k+r) - E_k^{(1)}] = 0.
\end{aligned}\right\} \quad (145)$$

These equations admit of a solution different from zero only in case the determinant of the coefficients vanishes. The corresponding secular equation for $E_k^{(1)}$ has $r+1$ real roots. If they are all different, the primary energy

[60] Cf. M. Born, "Vorlesungen über Atommechanik," Springer, Berlin, 1925, p. 287.

level E_k^0 is split into $r+1$ sublevels by the perturbing operator H_1 and the initial degeneracy is completely removed. In this case the $c^{(0)}$'s are uniquely determined by Eqs. (145) and the normalization condition. The $u_k^{(0)}$'s form a second complete orthogonal set of solutions of the unperturbed differential equation.[61]

As an opposite extreme to the case where all the roots of the secular equation are different comes the case where all are equal due to the fact that the diagonal elements of the $(r+1)$-rowed matrix $\{h_1(k,n)\}$ are all equal while the off-diagonal elements vanish. The initial degeneracy remains unaffected by the perturbation and the $c^{(0)}$'s from $c_{k,k}^{(0)}$ to $c_{k,k+r}^{(0)}$ are arbitary except for the normalization condition. It is then convenient to give one of them the value unity while equating the others to zero.

Intermediate cases in which the degeneracy is partly removed by the perturbation are of practical importance but need not receive detailed consideration here. Neither shall we take up the computation of the first order corrections to the characteristic functions for the case where E_k^0 is degenerate. For a discussion of this question, of the relation of the continuous spectrum to the perturbation theory, and of many other related questions the reader is referred to the original papers of Schrödinger.[61a]

2. The Variation Method and its Application to the Perturbation Problem.

A very different and instructive derivation of the fundamental equations (134) can be worked out from a variation principle equivalent to the Schrödinger wave equation and boundary condition. In its more general form this variational principle constitutes a wave-mechanical equivalent of Hamilton's principle in the classical dynamics. In a less general form it gives a valuable formulation of the characteristic value problem of Eqs. (15) and (37). The possibility of reducing the boundary value problems of the classical physics to calculus of variations form has long been recognized by mathematicians and has been of fundamental importance in the development of the general theory of these characteristic value problems. In the wave mechanics the variational method is as old as the differential equation method, for Schrödinger's first contribution to the subject begins with a deduction of the wave equation from an assumed variational principle.[62]

To set up the variational equivalent of the wave equation for a single particle in the general form (18) one may proceed as follows.[63] Let the wave equation be reduced to the simple form $L\psi=0$ by the introduction of the operator

$$L=H-\frac{h}{2\pi i}\frac{\partial}{\partial t}=-\frac{h^2}{8\pi^2\mu}\nabla^2+V(x,y,z)-\frac{h}{2\pi i}\frac{\partial}{\partial t}, \tag{146}$$

[61] The problem of Eqs. (145) is identical with that of the principal axis transformation of the quadratic form having the matrix $\{h_1(k,n)\}$ as explained in Sect. 4, 2. The orthogonality of the $u_k^{(0)}$'s is a simple corollary on the theory of such transformations.

[61a] E. Schrödinger, Ann. d. Physik (4) **80**, 437 (1926); **81**, 109 (1926); **83**, 556 (1927).

[62] E. Schrödinger, Ann. d. Physik (4) **79**, 361 (1926).

[63] Our discussion of this subject is a modification of that given by Sommerfeld in his "Ergänzungsband," Kap. I 9d.

and let M denote the conjugate operator $H+(h/2\pi i)\partial/\partial t$. By Green's theorem applied to any pair of suitably continuous functions $u(x,y,z)$, $v(x,y,z)$ defined throughout a volume G enclosed by a surface S

$$\iiint_G vLu\,d\tau = \iiint_G \left[\frac{h^2}{8\pi^2\mu}\left\{\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}+\frac{\partial u}{\partial z}\frac{\partial v}{\partial z}\right\}+uvV - \frac{h}{2\pi i}\cdot v\frac{\partial u}{\partial t}\right]d\tau - \iint_S v\frac{\partial u}{\partial n}dS. \tag{147}$$

$$\iiint_G uMv\,d\tau = \iiint_G \left[\frac{h^2}{8\pi^2\mu}\left\{\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}+\frac{\partial u}{\partial z}\frac{\partial v}{\partial z}\right\}+uvV + \frac{h}{2\pi i}\cdot u\frac{\partial v}{\partial t}\right]d\tau - \frac{h^2}{8\pi^2\mu}\iint_S u\frac{\partial v}{\partial n}dS. \tag{148}$$

Subtraction yields

$$\iiint_G [vLu-uMv]d\tau = -\frac{h}{2\pi i}\frac{\partial}{\partial t}\iiint_G uv\,d\tau + \frac{h^2}{8\pi^2\mu}\iint_S\left(u\frac{\partial v}{\partial n}-v\frac{\partial u}{\partial n}\right)dS. \tag{149}$$

Let v be the complex conjugate of u and let the region G be a sphere whose radius approaches infinity. If $\iiint_\infty uu^*d\tau$ exists, the surface integral will vanish in the limit. Mu^* is the complex conjugate of Lu and must vanish if the latter quantity does. Hence Eq. (149) may be used to prove the constancy of $\iiint_\infty uu^*d\tau$ for functions which are quadratically integrable and satisfy the differential equation (18).[64]

Consider now the integral K defined by

$$K=\int_{t_0}^{t_1}dt\iiint_G \psi^*L\psi\,dx\,dy\,dz. \tag{150}$$

Its first variation is

$$\delta K=\iiint_{G,T}[\delta\psi^*L\psi+\psi^*L(\delta\psi)]d\tau dt. \tag{151}$$

where T denotes the time interval $t_0<t<t_1$. Using the theorem of Eq. (149) with u and v set equal to $\delta\psi$ and ψ^* respectively, we obtain

$$\delta K=\iiiint_{G,T}[\delta\psi^*L\psi+\delta\psi M\psi^*]d\tau dt - \frac{h}{2\pi i}\iiint_G[(\psi^*\delta\psi)_{t_1}-(\psi^*\delta\psi)_{t_0}]d\tau + \frac{h^2}{8\pi^2\mu}\iiint_{S,T}\left[\delta\psi\frac{\partial\psi^*}{\partial n}-\psi^*\frac{\partial(\delta\psi)}{\partial n}\right]dS\,dt. \tag{152}$$

[64] Cf. foot-note on p. 170; also p. 193.

Then if δK vanishes for all variations in $\delta\psi$ which reduce the last two integrals of the above equation to zero, we may conclude that ψ satisfies the wave equation (18) $[L\psi=0]$ in the region G, T.

To prove the above statement it is convenient to separate ψ and $L\psi$ into real and imaginary parts as follows

$$\psi=\psi_1+i\psi_2, \qquad \psi^*=\psi_1-i\psi_2,$$
$$L\psi=\phi_1+i\phi_2, \qquad M\psi^*=\phi_1-i\phi_2.$$

Then Eq. (152) reduces to

$$\delta K=2\int\int\int\int_{G,T}[\delta\psi_1\phi_1+\delta\psi_2\phi_2]d\tau dt.$$

$\delta\psi_1$ and $\delta\psi_2$ are independent and arbitrary except at the boundaries of the space-time region G, T. Hence, if $\delta K=0$, ϕ_1, ϕ_2, and $\phi_1+i\phi_2$ must vanish. This proves the theorem.

Now let the region G take the form of a sphere whose radius approaches infinity. Let K approach a finite limit K_∞. If $\delta K_\infty=0$ for arbitrary variations in ψ which vanish at infinity and for which

$$\int\int\int_\infty[(\psi^*\delta\psi)_{t_1}-(\psi^*\delta\psi)_{t_0}]d\tau=0 \tag{154}$$

we may conclude that Eq. (18) holds over all space and through the time interval T. This theorem (I) is a wave-mechanical analogue of Hamilton's principle.

By means of Eq. (147) we may give the principle a slightly different form (II) in which K_∞ is replaced by the integral K'_∞ defined by

$$K_\infty'=\int_{t_0}^{t_1}dt\int\int\int_\infty\left\{\frac{h^2}{8\pi^2\mu}\left[\frac{\partial\psi^*}{\partial x}\frac{\partial\psi}{\partial x}+\frac{\partial\psi^*}{\partial y}\frac{\partial\psi}{\partial y}+\frac{\partial\psi^*}{\partial z}\frac{\partial\psi}{\partial z}\right]\right.$$
$$\left.+\psi^*\psi V-\frac{h}{2\pi i}\psi^*\frac{\partial\psi}{\partial t}\right\}d\tau. \tag{155}$$

A restricted form of the principle (III) is obtained if one seeks to fulfill the variation principle (I) by means of a monochromatic function

$$\psi=ue^{2\pi iEt/h}.$$

Then K_∞ can be replaced by

$$J=\int\int\int_\infty u^*(H-E)u\,dxdydz \tag{156}$$

where H is the Hamiltonian operator. The requirement that δJ vanish for all variations in u which vanish at infinity and for which $\delta E=0$ is equivalent to the requirement that u satisfy the wave equation (15) or that

$$Hu=Eu. \tag{157}$$

The variational principle (III) may be written

$$\delta J = \delta Q - E\delta \int_{\infty} u^* u d\tau = 0 \tag{158}$$

where

$$Q = \int_{\infty} u^* H u d\tau . \tag{159}$$

If E is fixed in the proper manner, Eq. (158) is the same as the pair of equations

$$\delta Q = 0 \; ; \; \delta N = \delta \int_{\infty} u^* u d\tau = 0 . \tag{160}$$

From Eqs. (156) and (157) it is clear that the desired characteristic value of E is given by

$$E = [Q/N]_m \; ; \quad N = \int_{\infty} u^* u d\tau . \tag{161}$$

if the subscript m be used to denote the "extremalized" value of the bracketed ratio. This last form of the variation principle (as stated in Eqs. (160) and (161)) we call (IV). It presupposes the existence of the integrals Q and N^*, and is therefore restricted to energy levels belonging to the discrete spectrum, but is of great importance in the approximate theoretical determination of these levels and of their wave functions.

The variational principles (III) and (IV) can be modified like (I) through the substitution of the quantity $\Lambda = (h^2/8\pi^2\mu)$ (grad u^* grad u) $+ u^* u V$ for $u^* H u$ in the integrands of the integrals to be varied. It is in these last forms that the variational principles are stated by Schrödinger[65] and Sommerfeld (l.c.). The extension of these principles to the n-body problem is obvious.

In the classical mechanics one of the most important uses of the variational principles is in deriving the differential equations of motion in generalized coordinates. To change from one coordinate system to another one has only to transform the integrand of the integral to be varied from the old system to the new and then write down Euler's equations for the minimization of the integral using the transformed integrand. The same use is made of the variational principles in the wave mechanics.[66]

After this rather lengthy introduction, we turn to the application of the variational method in the form (III) to the perturbation problem. To this end we use the Ritz direct method of attack on the variational problem.[67] Let $u_1^0, u_2^0, \cdots u_n^0, \cdots$ denote as before a complete normalized orthogonal system of discrete functions in terms of which we can expand any suitably continuous quadratically integrable comparison function u to be considered in connection with our variational problem. Let the expansion be

$$u = \sum_{n=1}^{\infty} c_n u_n^0 \tag{162}$$

[65] E. Schrödinger, l.c. and Ann. d. Physik **82,** 265 (1927). Cf. also W. Gordon, Zeits. f. Physik **40,** 117 (1926).

[66] E. Schrödinger, Ann. d. Physik (4) **79,** 734 (1926); V. Fock, Zeits. f. Physik **38,** 242 (1926).

the c_n's being arbitrary except for the fact that the sum of their squares must converge to a definite limit. Let Hu_n^0 have the expansion

$$Hu_n^0 = \sum_{m=1}^{\infty} H(n,m)u_m^0. \tag{163}$$

Then term by term integration of J converts it into the infinite quadratic form.

$$J = \sum_n \sum_m c_n c_m^* [H(n,m) - E\delta_{nm}]. \tag{164}$$

In order to minimize J we differentiate with respect to the real and imaginary parts of c_n and equate the derivatives to zero. By adding the two equations for each value of n one obtains the infinite set of equations (134, 5).[68]

The coefficients $c_1, c_2, \cdots c_n, \cdots$ may be regarded as an infinite set of complex coordinates which determine the function u. In the special case that these coefficients are real they may be interpreted as the orthogonal components of a vector in a space of infinitely many dimensions (function space).[69] In the more general case where the c's are complex it is still helpful to use a geometrical language based on the above-mentioned possibility to describe the characteristic value problem. As indicated in Eq. (138) the orthogonality of two functions U, V with components a_n and b_n, respectively, is equivalent to the statement that

$$\sum_n a_n b_n^* = 0.$$

As this is a generalization of the formula for the orthogonality of two vectors in three dimensions we may say that when two functions are orthogonal in the sense of Eq. (77) the corresponding vectors in function space are mutually perpendicular. Normalized functions are correlated with points on the unit hypersphere

$$N(c_n) \equiv \sum_n c_n c_n^* = 1.$$

If we express the integral $Q = \int_\infty u^* H u \, d\tau$ in terms of the components of u it becomes a homogeneous quadratic form and the equation

$$Q(c_n) = \sum_n \sum_m c_n c_m^* H(n,m) = E$$

represents an infinite family of similar "ellipsoids" in function space. In extremalizing J and Q we determine the directions in which the unit hypersphere is tangent to one of these ellipsoids, i.e., the principal axes of the family. The lengths of these principal axes for any fixed ellipsoid are inversely proportional to the square roots of the characteristic values $E_1, E_2, \ldots$. Hence the *lowest* characteristic value E_1 is the absolute minimum of Q for any point on the unit hypersphere or the absolute minimum of Q/N. The other values are minima with respect to displacements in certain directions and maxima with respect to others. It is possible, however, to set up a restricted variation problem for which any particular characteristic value, say E_τ, is an absolute minimum. To do so we must find all linearly independent characteristic functions $u_1, u_2, \cdots u_{\tau-1}$ for the energy levels below E_τ. Then E_τ is the absolute minimum of Q/N for comparison functions which are orthogonal to $u_1, u_2, \cdots u_{\tau-1}$.[70]

The process of minimizing or extremalizing J and Q and the location of a complete normalized orthogonal set of solutions u_n of Eq. (128) is the equivalent of performing the

[67] W. Ritz, J. reine angew. Math. **135**, 1 (1909); Courant-Hilbert, p. 157.

[68] For a more rigorous discussion with due emphasis on the dangers of the procedure see Courant-Hilbert, l.c.

[69] Cf. Courant-Hilbert, pp. 38, 39, 40.

[70] Cf. Courant-Hilbert, Kap. VI, § 1, 2.

principal axis transformation upon the quadratic forms $J(c_n)$ and $Q(c_n)$.[71] In the expansion

$$u_n = \sum_m \xi_{nm} u_m{}^0$$

for a normalized orthogonal set of u_n's the coefficients ξ_{nm} must constitute a normalized orthogonal solution of Eqs. (134). If we describe the comparison function u of Eq. (162) in terms of the u_n's instead of the $u_n{}^0$'s the expansion becomes

$$u = \sum_n d_n u_n$$

where

$$c_m = \sum_n \xi_{mn} d_n . \tag{165}$$

This equation describes an orthogonal linear transformation of coordinates in function space. If J is expressed in terms of the d's by means of this transformation it takes the normal form

$$J = \sum_n d_n d_n{}^* (E_n - E) . \tag{166}$$

By analogy with the analytic geometry of the conic sections we call this the "principal axis transformation". Such a geometrical interpretation is of course equivalent to saying that the matrices $\{H(n,m)\}$ and $\{H(n,m) - E\delta_{nm}\}$ of Q and J are reduced to diagonal form when referred to the correct characteristic functions of the problem in hand.

It seems probable that the variational method of deriving the perturbational equations (134) can be extended to cases where continuous as well as discontinuous spectra are involved but so far as the writer is aware the details have not been worked out.

The variation principle suggests an important method of attack on the problem of securing an approximate theoretical determination of the lower energy levels of a given system and their characteristic functions which does not involve the conventional expansion in powers of λ. In the case of the lowest energy level, for example, let an expression $u(q)$ be given, perhaps by the solution of one or more related problems, which constitutes a reasonable first approximation to the wave function desired. $u(q)$ can be generalized by the insertion of one or more parameters $\alpha_1, \alpha_2, \cdots \alpha\tau$ in order to give $u(q, \alpha_1, \cdots \alpha_\tau)$ maximum flexibility consistent with reasonable mathematical simplicity. It is then plausible to suppose that by a proper choice of the parameters $\alpha_1, \cdots \alpha_\tau$ we shall obtain a much better approximation than the original $u(q)$. To get the best values of the α's and also an approximate energy value one may insert $u(q, \alpha)$ into Eq. (159) and determine Q/N as a function of the α's. Minimization with respect to these parameters fixes the best values of the α's and at the same time locates the best energy values obtainable by this method. It is, of course, impossible to predict in advance the success of such a procedure, but the computed energy is sure to give an upper limit to the correct value (Cf. p. 81) and comparison with experimental data usually permits an estimation of the error after the completion of the calculation. In practice the uncertainty of this method is usually shared by the λ-series procedure owing to the excessive labor involved in working out the higher terms in the series.

The method just described can be applied to a few of the upper energy levels in case the wave functions have suitable symmetry characteristics.

[71] Cf. Madelung, "Math. Hilfsmittel" pp. 2, 3; Riemann-Weber I, Kap. 2, 2; Courant-Hilbert, Kap. I, § 3.

Thus if the lowest wave function is known to have a certain type of symmetry it may be possible to insure the orthogonality of this lowest function and a flexible comparison function $u(q, \alpha)$ by giving the latter a different type of symmetry. Then in accordance with the theory of p. 81 minimization of $Q(\alpha)/N(\alpha)$ yields an approximate energy value and wave function for the lowest of the energy levels having the corresponding type of symmetry.[72]

A quite distinct method of successive approximations has been developed by Brillouin who expands $\log \psi$ as a power series in h. The first approximation to the energy levels is then given by the Bohr theory with half-integral quantum numbers. Cf. Sect. 2, 4 and the references in footnote 29, p. 182.

3. Perturbations Involving the Time

The problem of perturbations involving the time has been treated by Schrödinger,[73] Dirac,[74] Born,[75] and Slater.[76] It is of fundamental importance for the discussion of the absorption, dispersion, and emission of radiation by matter—a problem to be considered in Section 5.

The most general form of wave equation derived up to this point is Eq. (38) in which V denotes a function of the coordinates only. We assume the possibility of modifying this equation, however, to include an interaction between the atom and a variable external electromagnetic field. Anticipating a result to be justified in the next section we write the modified equation in the symbolic form

$$\left(H-\frac{h}{2\pi i}\frac{\partial}{\partial t}\right)\psi=\left(H_0+H_1-\frac{h}{2\pi i}\frac{\partial}{\partial t}\right)\psi=0. \tag{167}$$

Here H_0 represents the unperturbed Hamiltonian operator of Eq. (120) or an equivalent in generalized coordinates. H_1 is a perturbing operator which in the simplest case is merely a potential V_1, depending on t as well as on the coordinates.

Following Born (l.c.) we assume that H_1 vanishes outside the time interval $0<t<T$. For negative values of t any permissible wave function must then reduce to the form

$$\psi=\sum_n c_n u_n e^{2\pi i E_n t/h}+\int_\epsilon^\infty c(E)u_E e^{2\pi i E t/h}dE \tag{168}$$

[72] The first application of the variation method to the characteristic value problem was the computation of the energy of normal He and normal Li^+ by Kellner (Zeits. f. Physik **44,** 91 and 110 (1927). The most conspicuous success is that attained by Hylleraas in his recent recomputation of the energy of normal He (Zeits. f. Physik **54,** 347 (1929)). The well-known computation of the energy of normal H_2 by Heitler and London, although carried through along lines closely paralleling the usual λ-series procedure for degenerate systems, is perhaps best justified as an example of the variation method. An excellent example of the application of the method to an excited state is contained in the paper on the B state of the H_2 molecule by Zener and Guillemin (Phys. Rev.—in press).

[73] E. Schrödinger, Ann. d. Physik (4) **81,** 109 (1926); **83,** 956 (1927).

[74] P. A. M. Dirac, Proc. Roy. Soc. **A112,** 661 (1926).

[75] M. Born, Zeits. f. Physk **40,** 109 (1926).

[76] J. C. Slater, Proc. Nat. Acad. Sci. **13,** 7 and 104 (1927).

where u_n and u_E are characteristic functions of the unperturbed problem (Eq. (127)) with the characteristic values E_n and E respectively. (We drop the superscript "0" for the unperturbed wave functions used in Sect 4, 1 as no other characteristic functions appear in our present discussion). In case the energy levels E_n are degenerate, we assume the u_n's form a complete normalized orthogonal set of characteristic functions.

As the expansion (168) is valid for any continuous function of the coordinates which behaves properly at infinity, it will hold for all values of t if the c's are made suitable functions of the time. As in the preceding articles we give formulas for the c's only in the idealized case where there is no continuous spectrum. The general case is discussed by Schrödinger. (l.c.).

Replacing H_1 by $\lambda F\{q, (h/2\pi i)\partial/\partial q, t\}$ in accordance with Born's notation and following the procedure of Sect. 4, 1 we find

$$Hu_n = \sum_m H(n,m)u_m = E_n u_n + \lambda \sum_m F_{nm}(t)u_m. \tag{169}$$

with

$$F_{nm}(t) = \int_\infty u_m^* F u_n d\tau$$

Introducing the expansions (168) and (169) into Eq. (167) one obtains the following infinite set of homogeneous differential equations for the c's

$$\frac{h}{2\pi i}\dot{c}_n = \lambda \sum_m c_m F_{mn}(t) e^{2\pi i (E_m - E_n)t/h} \quad n = 1, 2, 3, \cdots. \tag{170}$$

These relations take the place of Eqs. (134), (139) of Sect. 4, 1. For negative values of t the right hand members vanish as they should. The corresponding arbitrary constant initial values of the c's we indicate by the symbols c_n^0.

In order to retain our fundamental hypothesis regarding the physical interpretation of $\psi\psi^*$ it is necessary that any generalization of the wave equation such as Eq. (167) shall conform to the theorem of foot-note 11, p. 170 ($\partial/\partial t \int_\infty \psi\psi^* d\tau = 0$). In view of the discussion on p. 207 it follows that the basic operator H must satisfy the relation

$$\int_\infty [\psi^* H\psi - \psi H^* \psi^*] d\tau = 0. \tag{171}$$

If this important restriction is obeyed, we can reverse an argument due to Dirac (l.c.) to prove that the matrices $\{H(n,m)\}$ and $\{F_{nm}(t)\}$ are Hermitian. Then by hypothesis

$$\frac{\partial}{\partial t}\int_\infty \psi^*\psi d\tau = \frac{\partial}{\partial t}\left(\sum_n c_n c_n^*\right) = \sum_n (c_n^* \dot{c}_n + \dot{c}_n^* c_n) = 0. \tag{172}$$

By the aid of Eq. (170) we can throw the above equation into the form

$$\begin{aligned} &\frac{2\pi i\lambda}{h}\left\{\sum_{n,m}\left[c_n^* c_m F_{mn}(t) e^{2\pi i(E_m - E_n)t/h} - c_n c_m^* F_{mn}^*(t) e^{2\pi i(E_n - E_m)t/h}\right]\right\} \\ &= \frac{2\pi i\lambda}{h}\sum_{n,m} c_n^* c_m [F_{mn}(t) - F_{nm}^*(t)] e^{2\pi i(E_m - E_n)t/h} = 0. \end{aligned} \tag{173}$$

Since this quadratic form vanishes for arbitrary values of all the variables (subject only to the normalization condition) it follows that all the coefficients must vanish or that

$$F_{mn}(t) = F_{nm}^*(t). \tag{176}$$

This proves the theorem for $\{F_{nm}(t)\}$. As $\{H(n,m)\} = \{E_n\delta_{nm} + \lambda F_{nm}\}$ it is true also for $\{H(n,m)\}$.

The approximate solutions of Eqs. (170) given by other authors are contained in the following complete power series expansion due to Born (l.c.).

$$c_n = \sum_m c_m{}^0 f_{mn}(t). \tag{177}$$

Here

$$f_{mn}(t) = \sum_{p=0}^{\infty} f_{mn}{}^{(p)}(t)\lambda^p \tag{178}$$

with

$$\left.\begin{aligned} f_{mn}{}^{(p)}(t) &= \frac{2\pi i}{h}\int_0^t \sum_k f_{mk}{}^{(p-1)}(t')F_{kn}(t')dt' \ ; \\ f_{mn}{}^{(0)}(t) &= \delta_{mn}. \end{aligned}\right\} \tag{179}$$

For values of t greater than T (i.e., after the perturbation is over) Eq. (177) becomes

$$c_n = \sum_m c_m{}^0 f_{mn}(T). \tag{180}$$

The effect of the perturbation is thus expressible as a linear transformation of the c's with coefficients depending on the time. This transformation is *orthogonal*, for by Eq. (172)

$$\sum_n (c_n c_n{}^* - c_n{}^0 c_n{}^{0*}) = \sum_{m,k} c_m{}^0 c_k{}^{0*}\left[\left(\sum_n f_{mn} f_{kn}{}^*\right) - \delta_{mk}\right] = 0. \tag{181}$$

As the above statement is true for all values of the c's, we conclude that

$$\sum_n f_{mn} f_{kn}{}^* = \sum_n f_{mn} f_{nk} = \delta_{mk}. \tag{182}$$

Born bases his physical interpretation of these results on the postulate that an atomic system cannot occupy two energy levels at once. On this view the wave functions are to be used as the basis of a statistical mechanics dealing with ensembles of identical independent systems like the classical statistical mechanics of Gibbs.[77] Since $\sum c_n c_n{}^* = 1$, $c_n c_n{}^*$ is interpreted as the probability that an arbitrary system chosen from the ensemble is in the n^{th} state, or will occupy the n^{th} state if the perturbation is removed at the instant in question. It follows from Eq. (180) that if initially all the atoms are in the m^{th} state the final probability that any one of them will be in the n^{th} state is

$$\Phi_{mn} = f_{mn}(T) f_{mn}{}^*(T). \tag{183}$$

Born interprets the induced changes in the c's as quantum jumps and calls Φ_{mn} the transition probability for jumps from the n^{th} state to the m^{th}. The quantum jumps are not independent, however, since $c_n c_n{}^* = \left|\sum c_m{}^0 f_{mn}(T)\right|^2$.

[77] Cf. J. C. Slater, J. Franklin Inst. **207**, 449 (1929).

Born derives the Ehrenfest adiabatic theorem in the wave mechanics as a special consequence of Eqs. (177)–(179). He also treats perturbations involving the time in which the applied field is not removed after the time T. For details the reader is referred to the original paper (l.c.).

Foot-note added at reading of proof: All formulas in this article are based on the assumption of Eq. (17) that the sign of the exponent in the frequency factor of a monochromatic ψ function is *positive*. While this sign convention has been much used, notably in Sommerfeld's "Ergänzungsband," the author is inclined to the opinion that the opposite choice is preferable. The formulas for this alternative convention are obtained from those given above by replacing the symbol $i[(-1)^{1/2}]$ by $-i$ throughout. The operators used above are replaced by their conjugates which may be obtained direct from the appropriate classical formulas by the substitution $p_k \rightarrow +(h/2\pi i)\partial/\partial q_k$.

RECOMBINATION AND PHOTO-IONIZATION.

By F. L. Mohler
Bureau of Standards

Contents

PAGE
Introduction ... 216
Continuous absorption ... 217
X-ray absorption ... 217
Continuous absorption of alkali vapors ... 218
Line absorption ... 219
Photo-ionization ... 219
Recombination ... 221
Recombination spectra ... 221
Recombination in the afterglow ... 222
Measurements of recombination radiation ... 223
Summary of results ... 224

Introduction

EINSTEIN'S derivation of Planck's radiation law has introduced into physics a general method of deriving the numerical relation between the atomic constants involved in any two converse processes. Consequently experimental problems are facilitated inasmuch as there are two methods of approach available instead of one. The spontaneous recombination of an ion and an electron with emission of radiation and the converse process, ionization of an atom by absorption of radiation, are therefore two phases of the same problem.

A purely mechanical picture of the recombination process serves to visualize the atomic constants. If an electron approach a positive ion the path will be a hyperbola except as there is some dissipation of energy by radiation. The amount of radiation depends, according to classical theory, on the curvature of the path and speed which in turn depends on the distance between the ion and the extension of the original rectilinear path of the electron. Thus there is a limiting distance a determined by the electron speed v within which capture into a closed orbit occurs. For $v=0$ the initial path is a parabola and a becomes infinite since an infinitesimal dissipation of energy gives a closed orbit. The rate of recombination dN^+/dt is conveniently expressed in terms of the target area, $q(v)=\pi a^2$, since the number of effective collisions per cm^3 is

$$dN^+/dt = -vq(v)N^+N^- = -\alpha N^+N^- \qquad (1)$$

where the N's are concentrations of electrons and ions and α is called the recombination coefficient.

The energy radiated in recombination into a particular atomic level is the kinetic energy of the electron plus the negative energy, $h\nu_i$ of that level,

where ν_i is the frequency of the corresponding series limit. This radiation takes place in a single act with emission of a quantum

$$h\nu = h\nu_i + \tfrac{1}{2}mv^2. \tag{2}$$

Since the velocity will in general have a continuous range of values there will be a continuous spectrum extending toward higher frequencies beyond each series limit. The converse process is the absorption of light of frequency ν by an atom in the state of limiting frequency ν_i with consequent ionization of the atom by ejection of an electron with a velocity v. The variation of the absorption coefficient $k(\nu\nu_i)$ with ν in the continuous absorption beyond a series limit corresponds to the variation with electron speed of the capture area $q(v\nu_i)$ for capture into the level ν_i. If k is the atomic absorption coefficient (the absorption in one cm at a concentration of one atom per cm^3) then from the needle-quantum viewpoint it is a capture-area for quanta quite analogous to q for electrons.

The Einstein method of deriving the relation between these atomic functions is to set up the equations for the equilibrium in a black body inclosure at a temperature T, assuming that equilibrium holds for every pair of converse processes considered separately. There is then a necessary relation between $k(\nu\nu_i)$ and $q(v\nu_i)$ to satisfy the condition that both are purely atomic functions independent of temperature.[1]

$$q(v\nu_i)/k(\nu\nu_i) = 2ah^2\nu^2/m^2v^2c^2 \tag{3}$$

v and ν are related by Eq.(2) while a is the *a priori* probability of the state corresponding to ν_i (the number of magnetic levels) and c is the velocity of light. It is interesting that the right side of Eq.(3) is $2a$ times the ratio of the squares of the momenta of the quantum and the electron. As k is finite at the limit ν_i it follows from (3) that q must approach infinity for zero velocity as $1/v^2$ and the recombination coefficient $vq(v)$ will approach infinity as $1/v$.

Continuous Absorption

As concerns the choice of experimental attack, a measurement of the atomic absorption coefficients is the simplest in principle. In practice, experiments are limited to the comparatively few normal energy levels which lie in the near ultra-violet or in the region of penetrating x-rays. These levels are the 1^2S levels of the alkalies, the K levels of nearly all elements, the L levels of elements heavier than iron, and the M levels beyond tungsten.

X-ray Absorption.[2] Measurements of the atomic absorption coefficient beyond the K limit for all elements and all wave-lengths can be approximately represented by a single formula. The absorption by the K shell is

$$k_K = 1.91\times10^{-2}Z^4\lambda^3 \tag{4}$$

for wave-lengths less than the K limit and is zero for greater wave-lengths

[1] Milne, Phil. Mag. **47**, 209 (1925). Becker and Kramers also published derivations of this relation at about the same time.

[2] A. H. Compton, X-Rays and Electrons; Van Nostrand Co., N. Y.

where Z is the atomic number and λ is the wave-length in cm. Similarly

$$k_L = 0.25 \times 10^{-2} Z^4 \lambda^3. \tag{4a}$$

With the approximation $\nu/R = Z^2/n^2$ where n is 1 for K and 2 for L the absorption of any element at the limits λ_K and λ_L is

$$k_K = 1.58 \times 10^{-12} \lambda_K \tag{5}$$

$$k_L = 0.83 \times 10^{-12} \lambda_L. \tag{5a}$$

There is evidence that the exponent of λ in (4) and (4*a*) is not exactly 3 but the departure is never great.

Continuous Absorption by Alkali Vapors. Numerous observations show that there is continuous absorption beyond the principal series limits but quantitative measurements have been unsatisfactory. The metal is usually vaporized in steel tubes with water-cooled ends so that estimates of vapor density are very uncertain. Harrison[3] found that the continuous absorption of sodium extended throughout the ultra-violet with a maximum at the principal series limit 2412A and concluded that a large and unknown proportion of this absorption was of molecular origin. Ditchburn[4] attempted to separate molecular and atomic effects in potassium by measuring line and continuous absorption at various pressures but without much success. The absorption at high pressures increased continuously beyond the limit while at lower pressures it decreased and then increased in going from the limit, 2856A toward shorter wave-lengths. From the low pressure curves he estimated k as 10^{-19} at the limit.

Boeckner and the author[5] have made a measurement of the absorption of caesium vapor at a comparatively low pressure, 0.1 mm, under equilibrium conditions but the observed absorption coefficient of 4×10^{-19} is subject to great uncertainty because of the low value of the total absorption. Recently Trumpy[6] has published relative measurements of absorption in sodium vapor shown in Fig. 1. A background of molecular origin is evident

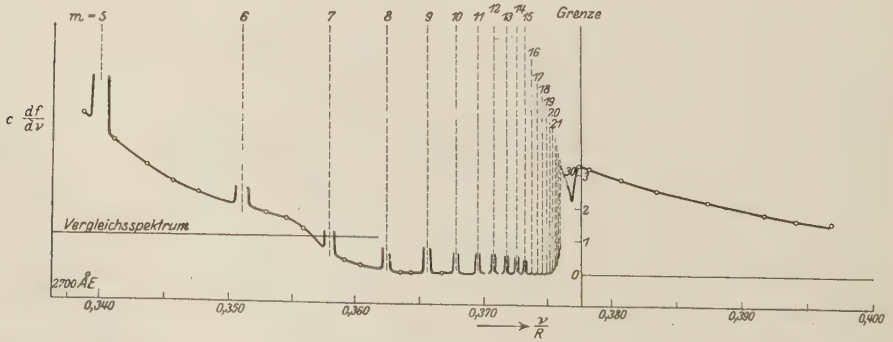


Fig. 1. Line and continuous absorption in sodium vapor (Trumpy). The wave-length scale extends from 2700A on the left to 2270A on the right with the limit at 2412.

[3] Harrison, Phys. Rev. **24**, 466 (1924).
[4] Ditchburn, Proc. Roy. Soc. **A117**, 486 (1928).
[5] Mohler and Boeckner, B. S. J. Res. **3**, 303 (1929).
[6] Trumpy, Zeits. f. Physik **47**, 804 (1928).

on the red side of the curve but is small near the limit so that he is justified in assuming that the absorption beyond the limit is atomic. This curve is much steeper than the λ^3 law of x-rays would give.

Line Absorption. Sugiura[7] has shown on theoretical grounds that in an absorption series the average value of the absorption coefficient over a definite frequency interval in the region of line absorption will approach in value the integral over an equal interval beyond the limit as we approach the limit. This offers a method of estimating the value of k at the limit which avoids the difficulty of molecular bands. The integral of $kd\nu$ across a line is difficult to measure and it is usually necessary to broaden the lines by high gas pressure or to use a thin layer at high vapor pressure to secure measurable magnitudes. Various results are given in the Handbuch der Physik[8] and in a recent paper by Waible[9] on series lines of caesium. If intervals $\Delta\nu$ are chosen equal to the average interval between lines to include one line in each interval, Waible's data for lines from the eighth to the fourteenth give

$$\int k d\nu/\Delta\nu = (0.48 \pm .01) \times 10^{-19} = k \text{ (limit)}. \qquad (6)$$

Measurements of most other observers show a similar (but less accurate) constancy though there are differences of as much as ten-fold between different observers. Other results are given in Table I.

TABLE I. *Absorption coefficients at the 1S limit.*

Method	Caesium	Rubidium	Potassium	Sodium
Photo-ionization	2.3×10^{-19}	1.1×10^{-19}		
Continuous absorption	$4\ \times10^{-19}$		1×10^{-19}	
Line absorption	0.5×10^{-19}	$5\ \times10^{-19}$		8×10^{-19}
	$5\ \times10^{-19}$			
Recombination	0.5×10^{-19}			

PHOTO-IONIZATION

A direct measurement of k can be obtained from a measurement of the current I produced in an ionization chamber of length l by a beam of radiation of known energy flux $J(\nu)$. Since k is small

$$k(\nu) = Ih\nu/eNpl J(\nu) \qquad (7)$$

where Np is the number of atoms per cm^3. The difficulty of avoiding surface effects has caused the development of other methods which give only relative values of k. One method[10] is to project a restricted jet of vapor through the ionization chamber; a second method[11] makes use of the neutralizing effect of positive ions on the negative space charge surrounding a hot wire cathode and the resulting electron current change to measure

[7] Sugiura, J. de Physique **8,** 113 (1927).
[8] Wolf and Herzfeld, Handbuch der Physik vol. 20, p. 616.
[9] Waible, Zeits. f. Physik **53,** 459 (1929).
[10] Lawrence, Phil. Mag. **50,** 345 (1925), and Williamson, Proc. Nat. Acad. **14,** 793 (1928).
[11] Foote and Mohler, Phys. Rev. **26,** 195 (1925).

photo-ionization. The space-charge method offers the advantage of extreme sensitivity since the ratio of electron current change to ion current may be as high as 10^5.

Boeckner and the author[12] have recently studied several of the alkali vapors by both the space-charge method and the direct method. The current in a long ionization chamber with a cathode carefully shielded from light gave the absolute value at a favorable wave-length while measurements of the space-charge effect gave relative sensitivity as a function of wave-length. Radiation flux was measured directly by thermopiles. Fig. 2 shows results in caesium and rubidium. The ordinates give relative sensitivity which is proportional to $k/h\nu$. The caesium curve which has a maximum at the limit 3184A has been displaced to coincide with the rubidium limit at 2967A. The absolute values of k at the limits are for caesium $(2.3 \pm 0.2) \times 10^{-19}$ and for rubidium 1.1×10^{-19}. There are minor discrepancies between observers. Little[13] gives a caesium sensitivity much higher at the high-frequency end, while Lawrence and Edlefsen[14] give both caesium and rubidium curves like the rubidium curve of Fig. 2. They, as well as other observers,

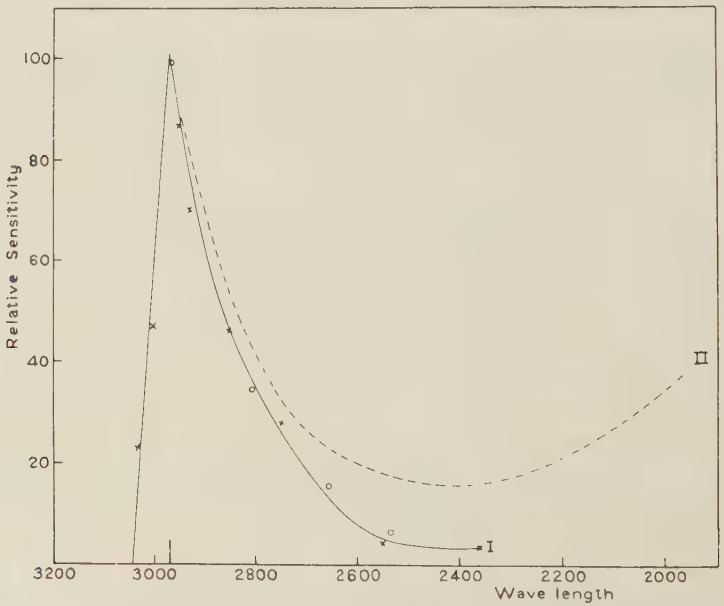


Fig. 2. Photoelectric sensitivity, I for rubidium and II for caesium. The caesium curve has been shifted to make the limit at 3184A coincide with the rubidium limit at 2967A.

obtain quite different results in potassium. The curve drops from the limit 2856A, as in Fig. 2, and then increases beyond 2700 to a maximum three times as high as the limit near 2340. It is commonly accepted that the increase beyond the limit in potassium and caesium can be ascribed to molecules

[12] Mohler and Boeckner, B. S. J. of Res. **3**, 303 (1929).

[13] Little, Phys. Rev. **30**, 109 (1927), and correction p. 963 (direct method).

[14] Lawrence and Edlefsen, Phys. Rev. **34**, 233 (1929) (space-charge method).

K_2 and Cs_2, though the shape of the curves does not change with pressure as might be expected on this hypothesis.

Recombination

With a notable exception measurements of the rate of disappearance of ions by recombination have been made at pressures comparable with an atmosphere under conditions such that the negative charges are heavy ions and not electrons. Results can be explained quantitatively on the basis of a theory of Thompson that recombination is effected by a three-body collision in which a neutral molecule takes up the excess energy;[15] so the experiments have little bearing on the problem of spontaneous recombination. Various lines of evidence indicate that under most low-pressure discharge conditions recombination in free space is negligible compared to loss of ions by diffusion to walls and electrodes.[16]

Recombination Spectra. The continuous spectra predicted by Eq.(2) give a safe criterion for the occurrence of spontaneous recombination and there are several striking instances of such spectra. Balasse[17] finds continuous spectra near the center of powerful induction ring discharges in caesium, potassium, mercury, and cadmium; while Herzberg[18] under similar conditions has photographed a continuous band beyond the Balmer series of hydrogen. Paschen[19] has observed recombination bands in a hollow cathode discharge in helium while the author has obtained such spectra in thermionic discharges at moderately high pressure and current.[20] These spectra are characterized by abnormally high intensity of higher series lines. The lines which are much broadened by high ion concentration merge into the continuous background which extends without any discontinuity through the limit and far to the violet side. The relative intensity in the series as well as the lines is abnormal. Thus in mercury and helium the triplet series are much stronger than the singlets. This may be ascribed to the higher *a priori* probability of triplet terms in accord with Eq. (3). Fig. 3 gives curves of intensity distribution of the continuous spectrum in caesium[20] and in helium.[21]

On the basis of considerations given in the introduction it is probable that low electron speed (small potential gradient) is more essential than high ion concentration for conspicuous recombination. An analysis of conditions in the thermionic discharges bears this out. The discharge tube consists of a hairpin cathode near the axis of a cylindrical anode and the voltage drop is within a thin bright sheath around the cathode. The recombination light in caesium appears as a diffuse yellow glow extending from the cathode toward

[15] Loeb, Kinetic Theory of Gases; McGraw Hill and Co.: Marshall; Phys. Rev. **34,** 618 (1929).

[16] Seeliger, Phys. Zeits. **30,** 329 (1929) in a very complete summary of recombination experiments includes a critique of negative results.

[17] Balasse, Comptes rendus **184,** 1002 and 1320 (1927).

[18] Herzberg, Ann. d. Physik **84,** 565 (1927).

[19] Paschen, Berl. Ber. p. 135 (1926).

[20] Mohler, Phys. Rev. **31,** 187 (1928).

[21] Mohler and Boeckner, Bur. Stds. J. of Res. **2,** 489 (1929).

the anode, and probe measurements[21] show that this region is characterized by high ion concentration (10^{12} to 10^{14} per cm^3) and very low electron speeds (0.1 to 0.4 volts). The space is slightly positive with respect to the anode so that slow electrons are entrapped to maintain a concentration nearly equal to that of the positive ions.[22]

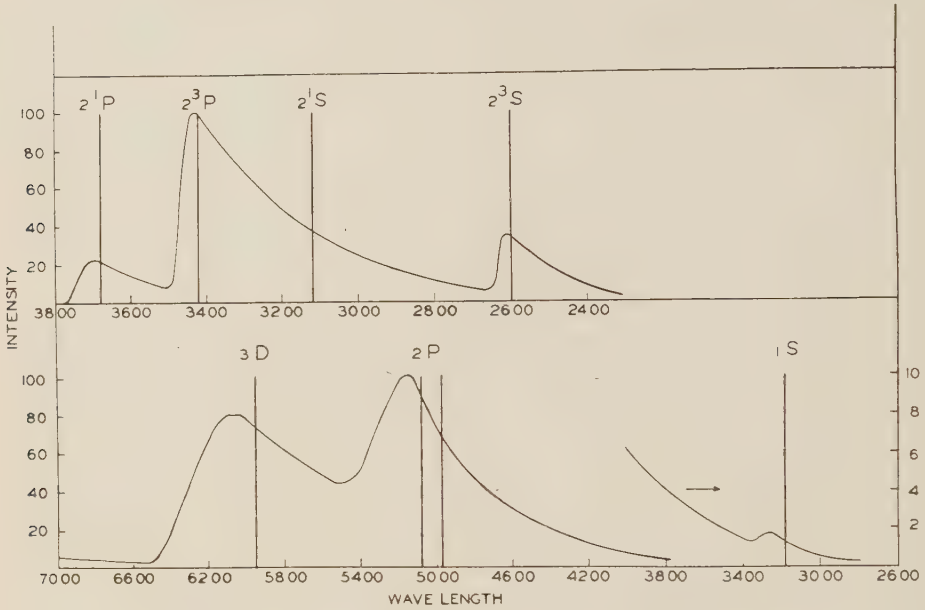


Fig. 3. Intensity distribution in the continuous emission spectrum of helium (upper curve) and caesium (lower curve).

Recombination in the Afterglow. Lord Rayleigh[23] has observed spectra indicative of recombination in the luminous vapor drawn from the discharge space in a mercury arc. Webb and Wang[24] find by probe measurements that the atomic radiation under these conditions is dependent on the presence of ions and very slow electrons, which have lost their speed in the course of diffusion. Miss Hayner[25] has studied the afterglow of mercury with a high speed sectored disk and commutator and finds that most lines fall to nearly zero intensity in 10^{-4} sec. after cutting off the voltage. But some lines (notably higher series lines of the diffuse triplets) then rise in intensity reaching a maximum about 3×10^{-4} sec. after the cutoff before finally fading. Her hypothesis that recombination is small at first and rapidly increases as the electrons lose their speed is supported by similar experiments of Kenty[26] in an argon discharge. By means of a suitable commutator he was able to make probe measurements at selected intervals during the dark period and afterglow. The electron speed dropped in about 2×10^{-4} sec. (the dark period) to

[22] Langmuir, Phys. Rev. **33,** 954 (1929).
[23] Rayleigh, Proc. Roy. Soc. **A108,** 262 (1925).
[24] Webb and Wang, Phys. Rev. **33,** 329 (1929).
[25] Hayner, Zeits. f. Physik **35,** 365 (1926).
[26] Kenty, Phys. Rev. **32,** 624 (1928).

0.4 volts and remained nearly the same for the duration of the afterglow (2×10^{-3} sec.). An accelerating voltage for electrons applied after the cutoff quenched the afterglow, though if the voltage was small enough the electron concentration was actually increased. This indicates that loss by recombination predominates over diffusion loss and his measurements of the rate of change of concentration give a value of the total rate of recombination which is certainly reliable as to order of magnitude. The electron concentration dropped from 7×10^{12} to 2×10^{12} in the interval from 2.5×10^{-4} to 2.5×10^{-3} sec. after the cutoff. Assuming $N^+=N^-$, Eq. (1) gives $\alpha=2\times10^{-10}$ for 0.4 volt electrons or $q=5\times10^{-18}$ cm^2.

Measurements of the Recombination Radiation.[21] The intensity $J(\nu)d\nu$ in a continuous band at a given frequency interval beyond the limit depends on the concentration of electrons of the corresponding speed, on the concentration of ions, and on the recombination coefficient for capture into the level concerned. Since the number of quanta emitted in unit volume equals the number of recombinations, we have corresponding to Eq. (1)

$$J(\nu)d\nu/h\nu=vq(v\nu_i)N^+N^-F(v)dv \tag{8}$$

where $F(v)dv$ gives the fraction of the electrons in a range corresponding to dv (Eq. 2).

Probe measurements show that the electron velocities in a thermionic discharge under recombination conditions have a strictly Maxwellian distribution about an average energy that can be accurately determined. Measurements of the spectral intensity of the radiation in the vicinity of the probe combined with the electrical measurements thus give relative values of $vq(v\nu_i)$ as a function of ν or v with a precision quite independent of uncertainties in N^+ and N^-. Figs. 4, 5, and 6 illustrate the experimental data and the computed curve for the band beyond the $2P$ limit of caesium. The two sets of measurements were made under quite different conditions of vapor pressure and ion concentration, yet the form of the vq curve remains the same. The curve of Fig. 6 is a plot of the equation

$$vq(v,2P)=k_1/\nu^2(\nu-\nu_i)^{1/2}. \tag{9}$$

Less accurate measurements on the band beyond the $3D$ limit of caesium fit the same curve. The maximum intensity at the $1S$ limit is about one percent of that at $2P$. At comparatively low intensities $J(\nu)$ is roughly proportional to the square of the discharge current or the square of the ion current to the probe. This is in accord with Eq. (8) if we assume that $N^+=N^-$; and on the basis of the ion current a rough estimate can be made of the numerical values of the concentrations. The data of Set II, Figs. 4 and 5 give the effective collision area for capture of 0.2 volt electrons into $2P$ as 6×10^{-21} cm^2, while for $1S$ the area is 1.5×10^{-23} cm^2. This is very much smaller than the total capture area found by Kenty for argon ions and 0.4 volt electrons ($q=5\times10^{-18}$ cm^2), and it seems probable that most of the recombination takes place into outer energy levels giving infra-red radiation. The great intensity of higher series lines supports this interpretation.

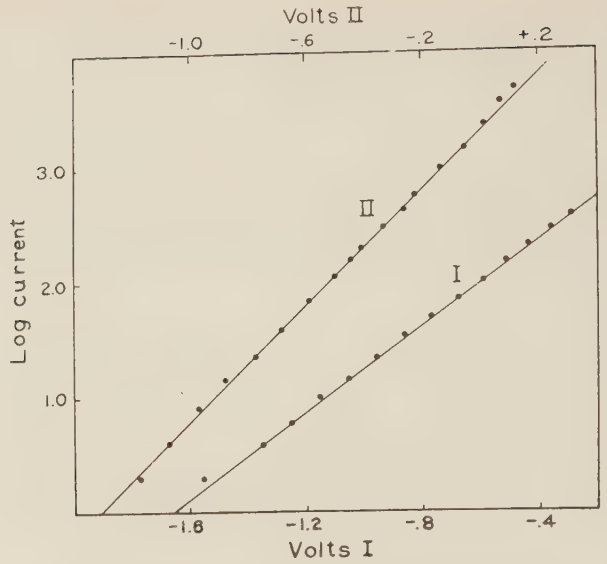


Fig. 4. Electron velocity distribution curves for two conditions of the caesium discharge. I gives an average electron energy of 0.345 volts; II gives 0.281 volts.

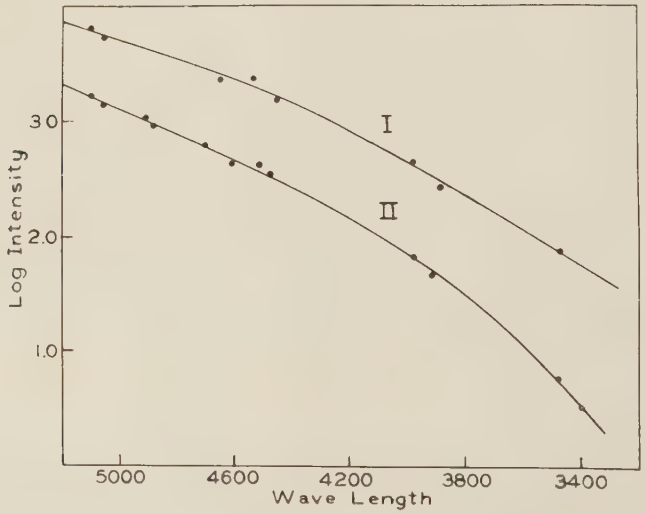


Fig. 5. Intensity distribution in the $2P$ band of caesium under conditions giving curves I and II of Fig. 4.

Summary of Results

On the basis of the x-ray law of absorption and Eq. (3), the law of variation of capture area is

$$q = c_1/\nu(\nu - \nu_i) = c_2/v v^2 \tag{10}$$

while from Eq. (9) based on intensity distribution in the caesium $2P$ band we have

$$q = c_1/\nu^2(\nu - \nu_i) = c_2/\nu^2 v^2 \tag{11}$$

where the c's are constants. The difference in form is small near the limit where the factor involving v is predominant. Conversely applying Eq. (3) to Eq. (11) we find that $k(\nu, 2P)$ is proportional to λ^4 instead of λ^3. It is important that the empirical Eq. (11) conforms to the requirement that k remains finite at the limit.

Fig. 7 gives the variation of k as derived from various sources. The curves for $1S$ levels of alkalies (potassium, as well as those plotted) fall close together near the limit and depart extremely from the x-ray law, while the absorption

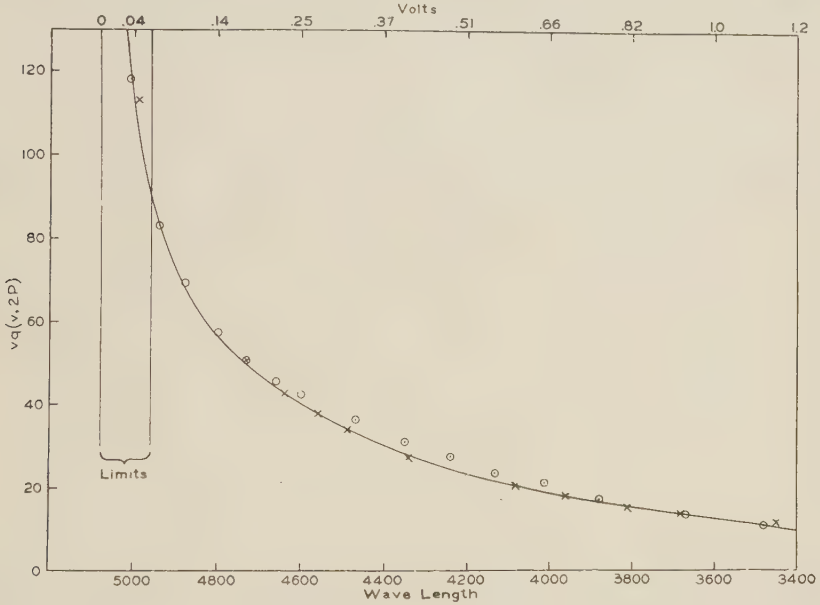


Fig. 6. Relative values of the coefficient for recombination into $2P$ from data of Figs. 4 and 5. Crosses refer to set I and circles to set II.

of caesium atoms in the $2P$ and $3D$ states is much like that of x-ray levels.

From the capture areas as given in the last section we derive by Eq. (3) that k for caesium atoms in the $2P$ state is 3×10^{-17} and for the normal $1S$ state is 0.5×10^{-19}. In view of the experimental uncertainties the latter value is in satisfactory agreement with other estimates. Table I summarizes the estimates of k for the alkalies covered in this paper. The values derived from direct measurement of photo-ionization are probably the most reliable.

Various theoretical derivations of the continuous absorption of hydrogenic atoms based on classical and correspondence principle considerations[27] lead to a λ^3 law of absorption. Wave mechanics derivations[28] show that the simple power law is only an approximation. Thus Oppenheimer finds for hydrogen in the 2_1 state that

$$k = 0.91\times10^{-12}\lambda^{2.9}/\lambda_i^{1.9}$$

[27] Kramers, Phil. Mag. **46**, 836 (1923).
[28] Oppenheimer, Zeits. f. Physik **41**, 268 (1927).

near the limit and that the exponent of λ approaches 3.5 for high frequency; while for the 2_2 state

$$k = 1.0 \times 10^{-12} \lambda^{3.4} / \lambda_i^{2.4} \tag{12a}$$

$$k \text{ (limit)} = 1.0 \times 10^{-12} \lambda \text{ (limit)} \tag{12b}$$

where wave-lengths are in cm. At the Balmer limit (3646A) $k(2_2) = 3.6 \times 10^{-17}$. This is of the same form and in close numerical agreement with the Eqs. (5) and (5a) for x-ray absorption. The formula for K absorption gives $k = 5.8 \times 10^{-17}$ at 3646, while the L series Eq. (5a) gives 3.0×10^{-17}. The estimate of k for caesium atoms in the $2P$ state was also 3×10^{-17} (12b would give 5×10^{-17} at $\lambda = 5000$).

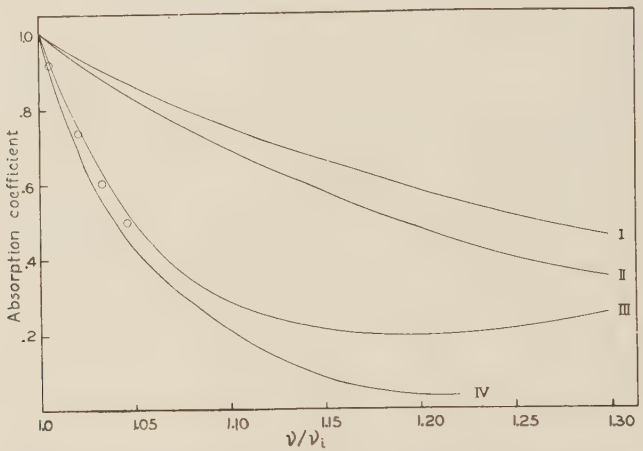


Fig. 7. Change of absorption coefficient with frequency. I from x-ray law, II from recombination into $2P$ of caesium, III from photo-ionization of caesium, IV from photo-ionization of rubidium. Circles from absorption of sodium (Fig. 1).

We have the striking result that the absorption curve for caesium atoms in the $2P$ and $3D$ states is closely hydrogenic both in magnitude and shape, while the $1S$ level departs widely from hydrogen in both respects. This anomaly can be partially understood. Dispersion and absorption theories (including wave mechanics) generally make use of the assumption that the integral of the absorption coefficient over all frequencies is for a single electron system equal to the absorption of a classical oscillator. Experiments on the alkalies show that, in contrast with hydrogenic atoms, nearly all this absorption is in the first line.[29] Consequently the absorption in the higher lines and beyond the limit must be correspondingly diminished. Sugiura[30] has carried out approximate computations by wave mechanics methods for the line absorption of sodium which are in satisfactory agreement with experimental values. Attempts by Hargreaves[31] and Trumpy[32] to extend such

[29] Wolf and Herzfeld, Handbuch der Physik, Vol. 20.
[30] Sugiura, Phil. Mag. **4**, 495 (1927).
[31] Hargreaves, Proc. Camb. Phil. Soc. **25**, 75 (1929).
[32] Trumpy, Zeits. f. Physik **54**, 372 (1929).

computations to the continuous absorption of lithium give curves which are not in mutual agreement and are quite unlike the experimental curves for heavier alkalies. The former gives a curve which reaches a maximum slightly beyond the limit and falls off roughly as λ^3, while the latter derives a curve which drops from the limit at a similar rate.

Experiments give a range of from 10^{-19} to over 10^{-17} for continuous absorption coefficients of atoms in various states, and theoretical considerations indicate that 5×10^{-17} is about the maximum for any limit in the visible or near ultra-violet. It follows that the area for capture into any level can not be much greater than that estimated for $2P$ of caesium (6×10^{-21} cm^2 for 0.2 volt electrons). This area is of a different order of magnitude from atomic dimensions (the cross section of the caesium ion is about 10^{-15} cm^2) so that the mechanical picture of the recombination process can only be accepted with reservations. The fact that under various discharge conditions recombination radiation becomes appreciable, in spite of these small magnitudes, is explained by the phenomenon of plasma formation.[22] The entrapping of slow electrons and consequent reduction of potential gradient in regions of high ion concentration brings about spontaneously a condition that is ideal for the experiment.

SPECTROSCOPY OF SOFT X-RAYS

By Thomas H. Osgood
University of Pittsburgh

I. Introduction

EVER since x-rays were recognised to be of the same nature as visible light, efforts have been made to explore, by photographic means, the unknown intervening region. Progress was rapid from both ends of the scale. Poineer work by Schumann[1] and Lyman[2] paved the way for the comprehensive experiments of Millikan[3] and his collaborators who, in 1921, extended the region of known optical spectra to wave-lengths as short as 136A. Meanwhile, the limit to which x-ray spectra from crystal gratings could be photographed was being pushed to longer wave-lengths, until, in 1925, Siegbahn and Thoraeus[4] successfully recorded lines at about 25A, —an important extension, as is easily realised by considering not wave-lengths, but frequencies. There remained a gap of about two and one-half octaves which defied exploration by standard methods. Further progress from the optical side was impossible on account of the extremely small reflecting power of gratings (used at normal incidence) for short waves; x-ray spectroscopists required crystals with very large grating spaces which were not readily obtainable; and in both cases were added the experimental difficulties of working in a high vacuum, for these x-radiations have very feeble penetrating power. Indirectly, spectrum lines and absorption edges in this intervening region had been detected by photoelectric methods,[5] but these were capable of no great accuracy in determining the positions of lines and could give, at the best, but vague information as to their intensity. It was not until a new technique was established that this last gap—the soft x-ray region, as it is called,—yielded satisfactory spectrum photographs. The soft x-ray region has no definite limits, merging into the extreme ultra-violet on the one hand, and into the domain of standard x-ray crystal spectroscopy on the other. It will be convenient, however, to restrict the spectra considered in this article to wave-lengths longer than about 15A. The experimental and theoretical basis of the new technique will be discussed in Section III.

II. Dauvillier's Experiments

No account of soft x-ray spectroscopy would be complete without reference to the work of Dauvillier,[6] who was the first to carry out a successful

[1] Schumann, Wien. Ber. **202,** 625 (1893).

[2] Lyman, Spectroscopy of the Extreme Ultra-Violet, Longmans, London, 1914.

[3] Millikan, Proc. Nat. Acad. Sci. **7,** 289 (1921). Millikan and Bowen, Phys. Rev. **23,** 1 (1924).

[4] Siegbahn and Thoraeus, J. O. S. A. **13,** 235 (1926). Thoraeus, Phil. Mag. **2,** 1007 (1926).

[5] Holweck, De la Lumière aux Rayons X, Conférences Rapports, 1927.

[6] Dauvillier, Jour. de Phys. et le Rad. **8,** 1 (1927).

exploration of this region by photographic means. Although his experimental methods have been superseded by more accurate and less difficult ones, his work presents many points of interest which make it worthy of consideration here. It was, essentially, an extension of the standard methods of x-ray crystal spectroscopy, one of the chief problems being the preparation of a crystal of sufficiently large grating space. Soft x-rays are very readily absorbed by matter, so that a thick crystal is unnecessary. Dauvillier's solution of the problem was to deposit a thin layer of melissic acid ($C_{29}H_{59}COOH$) only a few hundred molecules thick on a flat lead plate. Under these conditions the long-chain molecules arrange themselves in a regular fashion, somewhat like the pile of a carpet. (Compare, for example, the experiments of Hardy and Miss Doubleday[7] on lubrication.) The resulting structure has the properties of a crystal. In the particular case of melissic acid, the grating space was found to be 73.5A, and of its lead soap 87.5A, which should reflect spectrum lines up to a theoretical limit of about 175A, which, however, cannot be reached in practice.

The Schumann plates which were used were easily fogged by scattered light from the filament of the x-ray tube. To prevent this, a very thin screen of magnesium oxide was interposed in the path of the radiation. This was made by depositing the oxide from a magnesium filament on a thin film of collodion stretched on a wire frame. The screen, although practically opaque to visible light, transmitted about 30 percent of a radiation of wave-length 40A, the 70 percent absorption being due about equally to the collodion and the magnesium oxide. The thinnest foils of gold and aluminum which were opaque to light transmitted only a few percent of this soft x-radiation.

The measurement of his plates was difficult on account of the diffuseness of many of the lines (due in part to imperfections of the crystal); at the same time refraction of the rays in the acid film displaced the spectra by varying amounts almost impossible to estimate. The numerical results of Dauvillier's work, therefore, have no great value. He recorded, however, for the first time, a few lines of oxygen, carbon, boron and thorium between 20 and 120A. His papers also give some rough estimates of the absorbability of these soft radiations in thin foils. These are the only measurements of this kind which have been made for soft x-rays.

III. The Refractive Index of X-Rays and its Significance

Until a few years ago optical gratings in vacuum spectrographs were invariably used at nearly normal incidence. The new technique for the exploration of the soft x-ray region consists in allowing radiation to fall on the grating at angles of incidence which are greater than 80°, and usually almost 90°. Under these conditions, the soft x-rays are reflected and spectra are formed. We shall proceed to discuss the principles which underlie this modification of the regular procedure.

The experimental basis is to be found in a paper by A. H. Compton,[8]

[7] Hardy and Miss Doubleday, Proc. Roy. Soc. **A101,** 487 (1922).

[8] A. H. Compton, Phil. Mag. **45,** 1121 (1923).

in which that author describes experiments on the reflection of x-rays from plane surfaces of glass and silver. A beam of x-rays of a definite wave-length of the order of 1Å, selected by reflection from a crystal, was allowed to fall at a glancing angle of a few minutes of arc on a polished plate of glass, mounted on the turntable of an ionization spectrometer. The intensity of the reflected beam was measured in the usual way by an ionization chamber. By progressively varying the glancing angle and changing appropriately the position of the ionization chamber, a limiting angle was found at which the x-rays ceased to be reflected from the plane glass mirror. The experiments were carried out for more than one wave-length, and the significant facts emerged that the critical glancing angle increased with the wave-length and was different for different mirrors; in general, the more dense the substance of the mirror, the greater the critical angle.

Professor Puccianti,[9] of the University of Pisa, seems to have been the first to suggest that replacing the mirror by a plane optical grating and recording the reflected beam photographically should produce a diffraction pattern corresponding to the wave-length of the incident x-ray beam. His idea was tested by Carrara[9] who obtained, however, a negative result, though he verified the metallic reflection of x-rays. Unaware of Carrara's work, A. H. Compton and Doan,[10] a year later, carried out the same experiment; they were successful in recording a diffraction pattern in three orders, from a speculum grating. The diffraction and total reflection of x-rays were thus definitely established, and at the same time a valuable check was provided of the wave-length measurements of x-rays by crystal methods. It appears probable, indeed, that this method will, in a short time, become the most accurate at our disposal for the measurement of wave-lengths, and, indirectly, for the determination of Avogadro's and related constants.[10]

In A. H. Compton and Doan's work no data for wave-length calculations involved derived constants, such as the grating space of a crystal. A short time afterwards, in describing similar experiments, Thibaud[11] reported the important fact that glass gratings give greater intensity of reflection, less fogging of plates and sharper lines than those of speculum metal, as indeed Wood and Lyman[12] also showed in the extreme ultra-violet. This has been the general experience of later investigators.

Since x-rays in air (or in a vacuum) can be totally reflected from a polished surface, it follows that their refractive index μ is less than unity, but by a very small amount, since the critical glancing angles are small. Hence the velocity of an individual x-ray in (say) glass or speculum metal is greater than the velocity of light. There is, however, no contradiction of the special theory of relativity, since the group velocity remains less than that of light.

[9] Carrara, Il Nuovo Cimento **1**, 107 (1924).

[10] A. H. Compton and Doan, Proc. Nat. Acad. Sci. **11**, 598 (1925). Bearden, Amer. Phys. Soc., Washington Meeting, April 1929, Phys. Rev. **33**, 1088 (1929). Bäcklin, Nature, March 16, 1929.

[11] Thibaud, Comptes rendus, Jan., 1926. Thibaud, Jour. de Phys. et le Rad. **1**, 13 (1927).

[12] Wood and Lyman, Phil. Mag. **2**, 310 (1926).

The refractive index of crystals for x-rays has been calculated by Stenström[13] and by Larsson[14] from the deviations from Bragg's Law ($n\lambda = 2d \sin\theta$) in high orders of reflection; it has been measured more directly for crystals by B. Davis[15] and his collaborators; and in Siegbahn's[16] laboratory the direct deviation of x-rays by a glass prism has been observed. The result in all cases is a value for μ slightly less than unity. There is thus a wealth of concordant experimental evidence.

From the theoretical side this is well supported. The expression for the refractive index of x-rays of frequency ν may be written[17]

$$\delta = 1 - \mu = \frac{e^2}{2\pi m} \sum_1^N \frac{n_s}{(\nu^2 - \nu_s^2)}.$$

Here, since the right-hand member is essentially positive ($\nu > \nu_s$), the refractive index μ is less than unity, but by a small amount only, so that it is convenient to consider its difference δ from unity; e, m, are the charge and mass of an electron; N is the number of electrons per unit volume of the refracting substance, comprising several groups, each containing n_s electrons whose natural frequency is ν_s. The formula can be derived by purely classical methods, starting from Maxwell's equations. It involves the assumption that the magnetic permeability of matter for high frequencies (of the order of 10^{17} per second) is unity to a very close approximation. On theoretical grounds this is justifiable since our ideas of permeability involve the orientation of magnetic units (molecules, atoms, electrons) which would be impossible at such great frequencies. There is, however, a dearth of experimental confirmation.

In the case of the reflection of ordinary x-rays from glass, the frequency ν of the incident beam is large compared with the natural frequencies ν_s of any of the electron groups. Hence the formula may be simplified to

$$\delta = 1 - \mu = \frac{Ne^2}{2\pi m \nu^2}.$$

If θ_c be the critical glancing angle for total reflection, then

$$\sin\left(\tfrac{1}{2}\pi - \theta_c\right) = \cos\theta_c = \mu\ ;$$

or, since θ_c is small,

$$\sin\theta_c = [2(1-\mu)]^{1/2} = (2\delta)^{1/2}.$$

This relation affords a convenient method of comparing experimental results with theory. In spite of occasional slight uncertainties as to the exact chemical composition of the glass mirrors, theory and experiment show agree-

[13] See A. H. Compton, Ref. 8.

[14] Larsson, Zeits. f. Physik **41**, 507 (1927).

[15] v. Nardroff, Phys. Rev. **24**, 143 (1924). Hatley, Phys. Rev. **24**, 486 (1924). Davis and Slack, Phys. Rev. **27**, 18 (1926). Slack, Phys. Rev. **27**, 691 (1926).

[16] Larsson, Siegbahn and Waller, Naturwiss. **52**, 1212 (1924).

[17] Darwin, Phil. Mag. **27**, 315 (1914). Ewald, Ann. d. Physik **54**, 519 (1918). Ehrenberg and Mark, Zeits. f. Physik **38**, 129 (1926). Kallmann and Mark, Ann. d. Physik **82**, 585 (1927). A. H. Compton, "X-Rays, and Electrons," Chap. VII.

ment to within a few percent. It follows from the formula that δ is to a first approximation proportional to the square of the wave-length. Hence, wave-lengths of soft x-rays obtained by crystal methods cannot be expected to have a high accuracy. Thibaud and Soltan[18] have, in fact, shown that some of Dauvillier's results are in error by nearly 8 percent.

Since $\sin\theta_c$, is approximately proportional to λ, larger glancing angles can be employed for x-rays of longer wave-length. For $\theta = 10°$, all radiations softer than about 40A are reflected from a glass grating. In many cases larger glancing angles simplify the experimental procedure.

IV. The Use of Gratings at Large Angles of Incidence

In the soft x-ray region, as in optics, concave and plane gratings may be used for spectrum work. Each has its special advantages which recommend it for some particular investigation. The Rowland (concave) mounting requires that the slit, the center of the grating and the photographic plate lie on the circumference of a circle whose diameter is equal to the radius of curvature of the grating. This circle must be tangential to the grating, as in Fig. 1, where S is the slit, P the plate and G the concave grating. R is

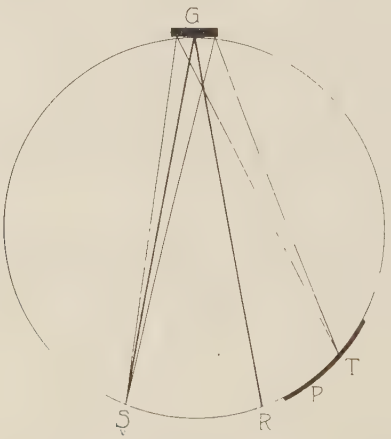


Fig. 1. Standard Rowland mounting for a concave grating.

the directly reflected beam. Under these conditions the spectra are in correct focus. For soft x-rays the glancing angle must be small; hence S and P must be moved along the circumference towards G. The resulting configuration is as shown[19] in Fig. 2, on a much larger scale than Fig. 1. The dotted arc represents part of the circle of reference. C is the source of radiation. If the grating is suitably adjusted an image of the slit, by direct reflection is formed at R, where $SG = GR$, and spectra are found between R and the outer end of the photographic plate, for example, at T. Since the angle of incidence is so large, the problem of housing even a two-metre grating in a vacuum spectrograph is not so formidable as might appear at first sight.

[18] Thibaud and Soltan, Jour. de Phys. et le Rad. **8,** 484 (1927).
[19] Osgood, Phys. Rev. **30,** 567 (1927).

This arrangement was used by Osgood[20] in the first successful exploration of the soft x-ray region from 40A to about 200A, overlapping considerably the extreme ultra-violet of Millikan.

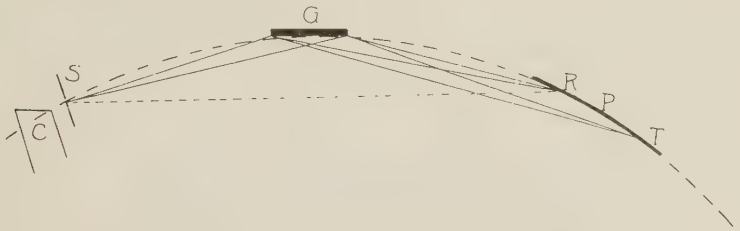


Fig. 2. Rowland mounting for large angles of incidence.

Almost at the same time, Thibaud,[21] who had already recorded lines at about 50A, reported x-ray spectra from a plane grating up to about 170A. These two researches of Thibaud and Osgood definitely closed the gap between x-ray and optical spectra, and illustrated the power of the new methods. The plane grating arrangement is explained[22] by Fig. 3. A pencil of radiation from an anticathode C is selected by parallel slits S_1, S_2 and

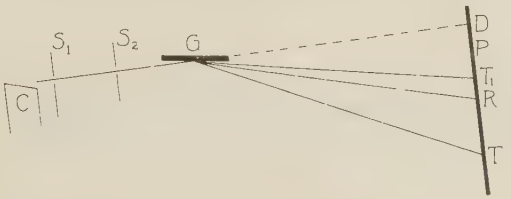


Fig. 3. Plane grating arrangement for large angles of incidence.

allowed to fall at a small glancing angle (perhaps 2°) on the grating at G. A reflected image is obtained on the plate at R, and spectra are found outside at T and inside at T_1. Since spectra cannot be formed behind the grating, those inside are limited. The distance GR is usually of the order of 50 cm. The slits S_1, S_2 may be made longer than the grating so that a reference line is produced at D while spectra are being recorded at T, T_1.

Up to the present time most investigators of the soft x-ray region have used plane gratings. They possess the following advantages over concave gratings: (1) measurements are accurate, approaching 0.1 percent; (2) spetrum lines are sharp[22] even with moderately wide slits; (3) adjustment is easy; (4) only a small grating is required. If a grating of the usual optical size were used, there would be uncertainty as to the exact part of it which was effective; that is, the point G, Fig. 3, would be indeterminate. For exact measurements, the position of this point must be known to a fraction

[20] Osgood, Nature, June 4, 1927.
[21] Thibaud, Comptes rendus, July 4, 1927.
[22] Thibaud, Jour. de Phys. et le Rad. **8,** 13 (1927).

of a millimetre. There are two ways of overcoming the difficulty. Either all the grating except a narrow strip 1 or 2 mm wide is masked by a thin film of collodion, or two photographs are taken simultaneously with plates at different distances from the grating.[23] The latter method is preferable, though, at the best, not more than a 10 mm strip of the grating is effective. Hence the intensity is limited. A concave grating, on the other hand, automatically focusses the spectra, and makes use of the whole surface of the grating. It is, unfortunately, difficult to adjust, and the wave-length measurements derived from it are lacking in accuracy.[19] There seems no reason to doubt, however, that it will develop into a valuable instrument for the study of soft x-rays as soon as a number of reference lines have been accurately measured by plane grating methods. The two types of grating are thus to be regarded as complementary instruments of research in the soft x-ray region. As regards dispersive power, for example, a good plane grating gives about 4A per mm at 100A, whereas a 2-metre concave grating with the same ruling gives 1A per mm if used at a glancing angle of roughly 10°. In neither case is the dispersion uniform. For a plane grating it is very great when the angle of diffraction is almost 90°, corresponding to a position on the photographic plate just outside the intersection of the planes of the grating and the plate, and decreases rapidly as the angle of diffraction decreases. (See Fig. 4 (c), 2).

The diagrams in Figs. 2 and 3 are schematic. For the detailed construction and adjustment of suitable vacuum spectrographs the reader must be referred to the original papers. In all cases wave-lengths are computed from a relation equivalent to

$$n\lambda = e(\cos\theta - \cos\phi),$$

which refers to a radiation of wave-length λ incident at a glancing angle θ on a grating whose constant is e, and diffracted in the n^{th} order at an angle ϕ. In the case of a concave grating, Fig. 2., the quantities necessary for the computation are: SR (straight line distance); RT, along the arc, i.e. along the plate; the grating space e; and the radius of curvature of the grating. For the plane grating, Fig. 3, the calculation requires, if the angle GDR is 90°: GD; DR; RT; and the grating space e.

In Fig. 4 are reproduced some typical soft x-ray spectrograms, taken with a plane grating. These I owe to the courtesy of Mr. Carl E. Howe, of Oberlin College, at present working at the University of Chicago. The strong lines D and R on all the photographs are the direct and reflected images of the slit system as in Fig. 3. X is due to diffraction of optical light (from the filament of the x-ray tube) past the straight edge of the grating. Spectra are formed on both sides of R. The interesting features of each photograph are enumerated below. All were taken with a plane glass grating of 6000 lines per cm; distance, plate to grating, about 25 cm; glancing angles from 1° to 5°.

[23] Thibaud, Jour. de Phys. et le Rad. 8, 447 (1927). Howe, Proc. Nat. Acad. Sci., March, 1929.

V. Sources of Radiation, etc.

In a vacuum spectrograph for soft x-ray work, the source of radiation is a Coolidge x-ray tube, attached directly to the body of the instrument so that no material screen intervenes between the anticathode and the grating. As it is operated at a comparatively low potential—of the order of 1 kv.—it may be made small, thus bringing the target close to the slit system, and

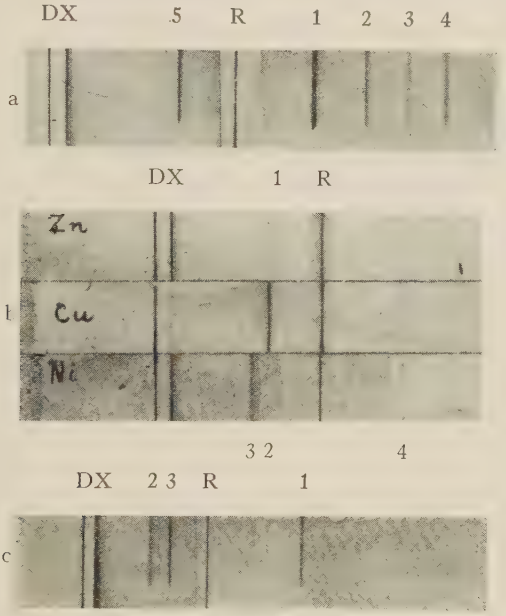


Fig. 4. (a) Anticathode, cobalt. Exposure 8 hours. 1, 2, 3, 4, Carbon K_α line, 44.6A, in four orders. This line is wide and diffuse. 5. Cobalt, $L_\alpha L_\beta$ overlapping; 16A.

(b) Anticathodes, zinc, copper, nickel. Exposures, 6, 1, 6 hours. 1. Zn, $L_\alpha L_\beta$ overlapping; 12A: 2. Cu, $L_\alpha L_\beta$ overlapping; 13A: 3. Ni, $L_\alpha L_\beta$ overlapping; 15A: 4. O, K_α; 24A.

(c) Anticathode, aluminum. Exposure, 12 hours. 1. Carbon K_α. 2. Al, K_α (strong); K_β (faint). 3. Si, $K_\alpha K_\beta$ overlapping. The dispersion of Al $K_\alpha K_\beta$ is much larger than of Si $K_\alpha K_\beta$ on account of the geometrical arrangement of the grating and plate, which produces non-uniform dispersion. The actual wave-length difference is only very slightly less for Si $K_\alpha K_\beta$ than for Al $K_\alpha K_\beta$. See Howe, Ref. 30.

thereby increasing the effective intensity of the x-rays. If the tube be made of metal and be water-cooled, it can carry large currents, 50–100 milliamperes, with safety. The brightness of the tungsten filament need cause no trouble if the anticathode be so shaped and orientated that it does not reflect light directly on to the slit of the spectrograph. Many tubes which fulfill these conditions are described in the literature.[24]

In the case of x-rays from an ordinary Coolidge tube, the fraction of the power input which reappears as x-rays is very small—about 0.1 percent. For x-rays of long wave-length, the value of this fraction is extremely uncertain. Thibaud,[23] from a consideration of the short exposures necessary

[24] Thoraeus, Phil. Mag. **1**, 312 (1926). Thibaud, Ref. 23.

to give good photographs, suggests that the conversion of the kinetic energy of electrons into radiation proceeds more economically in the soft x-ray than in the Roentgen region. On the other hand, Richardson[25] quotes the efficiency of soft x-ray excitation as being of the order of 10^{-6}, a figure based on the detection of these radiations by photoelectric methods. It is natural to expect that this efficiency should vary from element to element, and there is good evidence[26] that the variation is of a periodic character. Experiments of the photoelectric type are difficult to interpret since two transformations are involved; the primary electron-energy at the anticathode is partially converted into radiation which in turn causes the emission of electrons from another plate. There is no reason the problem should not be attacked by a spectrographic method. This should provide valuable information of a relative nature, at least. Whether absolute values of the efficiency of soft x-ray emission could be obtained is more doubtful, for it is well-nigh impossible to attain as good a vacuum in a spectrograph as in, say, the quartz apparatus employed by Richardson and his associates. The whole question is intimately bound up with the penetration of the cathode electrons through the surface layers of the target and the absorption by these layers of the radiation which is produced. A comparison of the absorption coefficients of slow-moving electrons and of their corresponding x-rays might lead to interesting results. Figures for the former, if we admit the mass-absorption law, are already available.

Early workers in this field used either Schumann plates, or x-ray plates smeared with a minute amount of fluorescent oil. With improved technique and more powerful sources of radiation, it has been found that ordinary untreated x-ray plates serve admirably, giving clear and sharp pictures. The first attempts at soft x-ray photography were made with speculum gratings, and met with varying success. The general characteristic of these gratings is to produce plates which are partially fogged, as though the grating scattered, rather than reflected, the radiation. As in optics, some gratings give better results than others apparently similar. Glass gratings give much better photographs; a light ruling seems to produce as intense spectra as a heavy ruling. It is rather remarkable that a grating, accurate perhaps to one-tenth of a wave-length of visible light, should reflect and diffract so well these short wave-length radiations.

VI. Interpretation of Soft X-Ray Spectra

In the early infancy of soft x-ray spectroscopy, there was little ordered investigation; the difficulty was to obtain spectra at all; subsequent improvements in technique have now reduced this almost to a routine performance; and in the last eighteen months a few systematic investigations have provided enough material for a study of the regularities in the spectra. One of the chief interests lies in the transition from x-ray to optical spectra. The regularity in x-ray spectra, first demonstrated by Moseley[27] for the *K*

[25] Richardson, Proc. Roy. Soc. **A119,** 538 (1928).

[26] Richardson and Robertson, Proc. Roy. Soc. **A115,** 280 (1927).

[27] Moseley, Phil. Mag. **26,** 1024, (1913); **27,** 703 (1914).

series, is that the square root of the wave-number of a particular line varies linearly with the atomic number of the element by which it is emitted. This may be expressed as

$$(\nu/R)^{1/2} = k(Z - k_1),$$

where ν is the wave-number of the line, R is Rydberg's constant, Z the atomic number, and k, k_1 are constants depending on the series (K, L, M, etc.) to which the line belongs and on the particular line of that series which is being considered. If the small relativity correction be ignored, this law represents faithfully all experimental wave-length observations, except for light elements. For the K series, the formula fails only in the case of hydrogen; for the L series, there is a break at neon ($Z = 10$); for the M series, at argon ($Z = 18$).[28] For lighter elements our knowledge is fragmentary and approximate. It is here that the importance of soft x-ray spectra is apparent.

Söderman[29] has recently published for soft x-rays the wave-lengths of the K_α lines of the elements from Be to Mg (excepting Ne), using the K_α line of Al as a standard, and deriving therefrom the constants of his spectrograph. His measurements overlap and extend the range covered by the earlier ones of Thibaud and Soltan,[18] and of Siegbahn and Thoreaus; among all three there is good agreement. The $(\nu/R)^{1/2}$ values for these K_α lines are plotted in Fig. 5 as a function of the atomic number of the corresponding elements. The curve has been completed by the points for

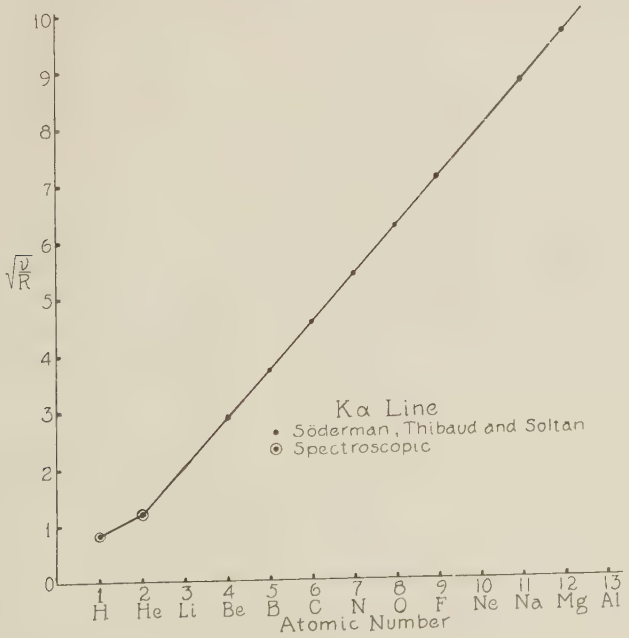


Fig. 5.

[28] Kossel, Zeits. f. Physik **2**, 470 (1920).
[29] Söderman, Zeits. f. Physik **52**, 795 (1929).

hydrogen and helium obtained from spectroscopic data. As predicted by Kossel, and in accordance with Moseley's law, the graph is straight as far as He, where the K shell of electrons is first completed. For hydrogen it is not complete, and the deviation of the curve from linearity is explained. Spectra of the L radiations of most of the elements from Ca(20) to Zn(30). have been photographed by Howe,[30] thus checking and amplifying the results already obtained by Thoraeus[4] by crystal methods. As is to be expected, their frequencies show a regular progression from one element to the next. The appearance of new lines in the L series as the atomic number increases indicates the building up of new groups of electrons in the atom, for, of course, an x-ray line cannot be emitted unless there is present in that atom at least one electron in the appropriate initial "stationary state." The evidence from observations of this nature by Thoraeus and Howe lends strong support to the accepted[31] grouping of the external electrons, at least as far as the low energy levels are concerned. The regular doublets of the N series (c. 50A) have been measured with some accuracy by Thibaud and Soltan[18] for the elements Ta (73), W (74), Pt (78) and Au (79). Apart from all these, our present knowledge of x-ray wave-lengths greater than 25A is extremely fragmentary.

The energies corresponding to the x-ray lines provide information only as to the relative amounts of energy associated with electron groups in the atom. The absolute values are derived from absorption discontinuities. These can be recorded with comparative ease in the Roentgen region, but in photographs of the longer wave-length region they are, in general, conspicuously absent. There are two reasons for this: (1) an x-ray tube operated at a low voltage appears to emit very little continuous radiation; (2) the radiation has very slight penetrating power. In spite of (2), one might expect to find on a photograph absorption edges characteristic of the material of the emulsion of the plate. There has been, however, one success. Using an anticathode of high atomic weight, Thibaud[32] demonstrated the emission of a continuous spectrum between 15A and 250A. In a metal spectrograph such as he used there are always present small quantities of carbon (vapors of grease, etc.), oxygen and nitrogen. These produced K absorption bands in several orders with sharp edges on the high-frequency side, so that the plate had a fluted appearance. The edges had wave-lengths as follows: C, 43.5A; N_2, 31.1A; O_2, 23.5A, as compared with the corresponding K_α lines, 44.9, 31.8, 23.8A. These figures permit an evaluation of the absolute energy of the $L_{II, III}$ levels in these atoms, since the sum of the energies of the K_α line and the $L_{II, III}$ level amounts to the energy of the K absorption edge. We are concerned here, however, with the differences between approximately equal numbers, as the energy of this level for these light elements is only a few electron volts. For elements heavier than Ne (10) the Moseley graph for the $L_{II, III}$ level is linear, as is to be expected (Fig.

[30] Howe, Amer. Phys. Soc., Washington Meeting, April 1929, Phys. Rev. **33**, 1088 (1929).
[31] Stoner, Phil. Mag. **48**, 419 (1924).
[32] Thibaud, Nature, March 3, 1928.

6.) For lighter elements, the available measurements indicate irregularities and a change in the slope of the graph, but their accuracy and completeness are scarcely adequate to determine definitely the nature of these irregularities. On theoretical grounds a break in the curve should occur at Ne (10). Precision measurements of lines and of absorption edges throughout this region are urgently required.

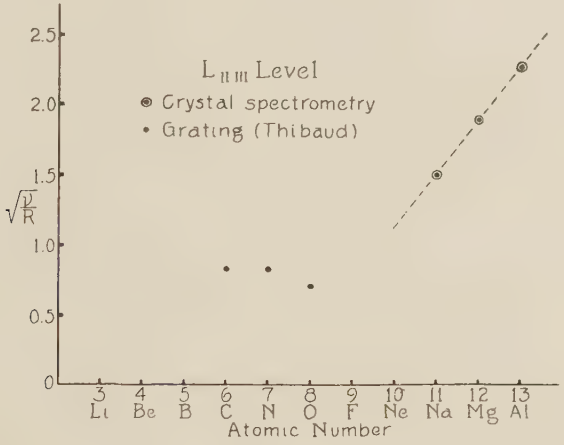


Fig. 6.

Already a beginning has been made in the accurate measurement of the carbon K_α line. This appears strongly on all plates, and is therefore a convenient reference line, although its diffuseness (about 1A wide) is a disadvantage. Thibaud[21] first gave its wave-length as 44.9A, with a possible error of two or three tenths of an Angstrom. A year later, Weatherby[33] arrived at the result 45.4 ± 0.2A but this appears to be much too high. In the summer of 1927 the writer, believing that an accuracy of 0.1 percent could be attained, found[34] the value 44.52 ± 0.05A, but unfortunately his experiments had to be discontinued before the data were ample enough to justify publication. They were resumed by Howe,[35] who recently completed a careful measurement of this line, giving as his final value 44.60 ± 0.04A. This seems to be the most reliable determination up to the present time. It agrees well with Söderman's[29] figure of 44.70 ± 0.09A. Other measurements have been made by Hunt[36] (with a grating) and Dauvillier[6] (with a crystal) but their probable error is large.

VII. Conclusion

This review has been concerned with x-ray spectroscopy by grating methods. There is a mass of information, obtained from crystal spectrom-

[33] Weatherby, Phys. Rev. **32,** 707 (1928).
[34] Osgood, Dissertation, Chicago, 1927.
[35] Howe, Ref. 23.
[36] Hunt, Phys. Rev. **30,** 227 (1927).

etry, which overlaps the beginning of the soft x-ray region, but this has been referred to only in so far as it has a direct bearing on the immediate subject. Interesting problems arise also in connection with the refractive index of soft x-rays. For these, μ is less than unity, while for the ultra-violet, the value is much greater than unity. Already observations have shown that in the Roentgen region there occur phenomena analogous to anomalous dispersion in optics. It is not inapt to quote the words of Lorentz, who, with Drude, developed the theory which predicts the total reflection of x-rays. "It is remarkably interesting that with x-rays, one finds near an absorption edge phenomena similar to those which, in classical optics, occur near an absorption band. There is, however, a profound difference between these two cases, as the absorption edge does not correspond to a frequency which actually exists in the particles." Up to the present time much of the soft x-ray work has aimed at verifying or improving the accuracy of crystal measurements, which, for long wave-lengths, are complicated by refraction and uncertainty as to the exact grating space of crystals. Comparatively few observations have been made for very long wave-lengths. This is natural. To provide an adequate and unequivocal interpretation of results, progress must be gradual, advancing by small steps from known to unknown territory. We may confidently look forward to obtaining from the spectroscopy of soft x-rays a complete understanding of the transition from x-ray to optical spectra, and accurate measurements of the absolute values of the low energy levels in the atom, which will, at the same time, show how the extra-nuclear structure is altered as the atomic number changes. This change is periodic. Harder x-ray spectra, considered as a function of atomic number, give no indication of periodicity, since the shells of electrons near the nucleus are little affected by changes in the nuclear charge.

Note added in proof: Since this article was written, Bazzoni, Faust and Weatherby (Nature, May 11, 1929, p. 717) have succeeded in resolving the carbon K_α line into four components. This explains the divergent values obtained earlier for the wave-length of this line, for the prominence of any particular component appears to depend on the potential applied to the x-ray tube.

ERRATA

Probable Values of General Physical Constants

RAYMOND T. BIRGE

Physical Review Supplement **1**, July, 1929

Page 27, line 8, read cm^3 for cc

Page 62, near top, read $h^2 = 42.86_{32}$ for 42.86_{42}

Page 63, last constant, read $a_0' = 0.5281_{69}$ for 0.5284_{69}

There are no typographical errors on the cards sent out in connection with this article, except that the final superscript 2 in the Schroedinger constant for the electron did not print.

REVIEWS
OF
MODERN PHYSICS

(Formerly the Physical Review Supplement)

CONDUCTED BY THE
AMERICAN PHYSICAL SOCIETY

BOARD OF EDITORS

JOHN T. TATE, University of Minnesota

A. H. COMPTON	K. T. COMPTON	K. K. DARROW
E. C. KEMBLE	C. E. MENDENHALL	D. L. WEBSTER

VOLUME 2

1930

CONTENTS

JANUARY, 1930

General Principles of Quantum Mechanics. Part II – E. C. KEMBLE AND E. L. HILL 1
Errata to Part I, General Principles of Quantum Mechanics – – – – – 59
Interpretation of Band Spectra. Parts I, IIa, IIb – ROBERT S. MULLIKEN 60
X-Ray Diffraction in Liquids – – – – – – – – – G. W. STEWART 116

APRIL, 1930

Electrical Discharges in Gases. Part I. Survey of Fundamental Processes – – – – – – – – – – KARL T. COMPTON AND IRVING LANGMUIR 123

JULY, 1930

Relativity and Electrodynamics – – – – – – – – – W. F. G. SWANN 243
Application of Group Theory to the Quantum Dynamics of Monatomic Systems – – – – – – – – – – – – – – – – CARL ECKART 305

OCTOBER, 1930

Thermionic Emission – – – – – – – – – – – – – SAUL DUSHMAN 381
Magnetism in Discontinuous Media – – – – – – – L. W. MCKEEHAN 477
Interpretation of Band Spectra, Parts I, IIa, IIb. Additions and Corrections – – – – – – – – – – – – – – – ROBERT S. MULLIKEN 506

Volume 2 January, 1930 Number 1

REVIEWS OF MODERN PHYSICS

THE GENERAL PRINCIPLES OF QUANTUM MECHANICS. PART II

By E. C. Kemble and E. L. Hill*
Harvard University

SECTION 5. THE INTERACTION OF MATTER AND RADIATION

1. General Introduction

As the primary source of experimental evidence regarding the structure of the atom is the spectroscope, much of the mathematical structure developed in the preceding sections must hang in the air until we have constructed a quantum theory of the emission, absorption, and scattering of radiation by atoms. The desired theory should have a formal similarity to the elementary wave mechanics of Part I and should degenerate in the limit when h is made negligibly small into the classical electromagnetic theory of the emission and absorption of light. It must take into account the forces acting on the electrons due to the external radiation field, the modification of the radiation field due to the motion of the system, and finally, the damping influence of "radiation resistance." To be completely satisfactory the theory should also take into account the effect of the finite velocity of electromagnetic waves on the interaction of the different electrons in the atomic system and should satisfy the demands of the theory of relativity. The newly formulated Heisenberg-Pauli field theory goes a long way toward the solution of this great problem[1] but it would be quite impossible to introduce the reader to the mysteries of this theory with the mathematical tools developed in Part I of this report. Progress toward carrying out the above program has been made by successive approximations and we here content ourselves with reporting the initial steps in the investigation.

In the usual classical discussion of absorption, dispersion, and scattering, the forces acting on the electrons are somewhat arbitrarily divided into three classes. First, there are the forces internal to the atom, whose operation determines its unperturbed state. According to our present ideas of

* National Research Fellow.

[1] W. Heisenberg and W. Pauli, Zeits. f. Physik **56**, 1 (1929).

atomic structure these forces are to be identified with the electrostatic forces which act between electron and nucleus and between electron and electron. Second, there are the forces due to the external radiation field, and third, the retarding force on each electron due to its own radiation.[2] The elementary quantum theory of Part I takes cognizance only of the first class of forces. It can be extended without difficulty to include forces of the second class, but the forces of the third class, difficult to deal with even classically, are particularly intractable from the standpoint of wave mechanics. Radiation resistance forces are small, however, in the case of an atom near to its normal state and by neglecting them we may still hope to obtain a description of the early stages of the absorption process. The problem of dispersion can also be treated without considering radiation resistance if means are at hand for calculating the electromagnetic field due to a given electronic motion.

As a first step toward the formulation of an approximate quantum theory of absorption and dispersion a wave equation analogous to Eqs. (I,18) and (I, 38) must be set up to describe the behavior of the electrons in an arbitrary external electromagnetic field.

In view of the success of the strictly corpuscular theory of the Compton effect, it might seem that any attempt to account for the interaction of matter and radiation in terms of a continuous electromagnetic field is foredoomed to failure. We have already seen, however, that the field theory of diffraction can be interpreted as a statistical description of the scattering of photons and electrons by apertures. It is such a statistical description of the interaction of matter and radiation that we seek. Moreover, the continuous potential energy function V of the Schrödinger Eqs. (I, 18) and (I, 38) is in practice (Cf. Kepler problem, Section 2, 6) of electrostatic origin and if we admit such a continuous field to our theory we can hardly stop short of admitting the whole machinery of Maxwell. The Heisenberg-Pauli field theory provides a scheme for the modification of this theory, but we shall provisionally content ourselves with an attempt to graft the electromagnetic field of Maxwell unchanged to the matter waves of Schrödinger.

2. The Wave Equation for Systems of Electrons in an External Radiation Field

The principle of least action in the Newtonian dynamics is restricted to cases in which the energy is conserved.[3] Hence the simple method of setting up the wave equation sketched in Part I, Section 1 cannot be used for a system of electrons in a variable external field. Schrödinger himself, however, based his "derivation" of the wave equation (I, 18) on the Hamilton-Jacobi partial differential equation which is not subject to the above restriction.[4] His point of view permits us to set up and justify a generalization of Eqs. (I, 18) and (I, 38) appropriate to the problem in hand.

We have already shown that in the case of a constant electrostatic field the wave equation [(18) or (38), Part I] is obtained from the classical

[2] With these we may include the damping effects of collisions which play such an important part in the production of absorption in gases.

[3] That is, cases in which there is a Hamiltonian function $H(q, p)$ which does not contain the time explicitly.

[4] E. Schrödinger, Ann. d. Physik (4) **79**, 361 and 489 (1926).

Hamiltonian function $H(q, p)$ if we first form the operator $H(q, -\partial/\partial q)$ by the substitution $p_k \rightarrow -(h/2\pi i)\,\partial/\partial q_k$ and then use it to set up the relation

$$H\left(q_k, -\frac{\partial}{\partial q_k}\right)\psi = \frac{h}{2\pi i}\frac{\partial \psi}{\partial t}. \tag{1}$$

The parallelism between this second order first degree equation for ψ and the first order second degree Hamilton-Jacobi equation for the action function

$$H\left(q_k, \frac{\partial W}{\partial q_k}\right) = -\frac{\partial W}{\partial t} \tag{2}$$

W is obvious. A complete solution of the latter equation is a function of the coordinates, say $q_1, q_2, \cdots q_{3n}$, of the time, and of $3n$ constants of integration $\alpha_1, \cdots, \alpha_{3n}$, which represents a moving family of surfaces in configuration space $(q_1, \cdots, q_{3n})$ related to the possible mechanical orbits of the classical dynamics as the wave-fronts of geometrical optics are related to light rays. As wave fronts are surfaces of constant phase it is therefore not suprising to find that the classical action function is simply related to the phase of the ψ waves. If we change the independent variable in Eq. (I, 38) by means of the substitution

$$\psi = e^{-2\pi i W'/h} \tag{3}$$

we obtain the differential equation,

$$\sum_k \frac{1}{2\mu_k}\left(\frac{\partial W'}{\partial x_k}\right)^2 + V + \frac{\partial W'}{\partial t} = \frac{h}{4\pi i}\sum_k \frac{1}{\mu_k}\frac{\partial^2 W'}{\partial x_k^2}. \tag{4}$$

If the second derivatives on the right hand side of Eq. (4) are neglected, it becomes identical with the Hamilton-Jacobi equation (2) for the problem in hand. This approximation is equivalent to setting Planck's constant h equal to zero, and is valid in the region for which the classical mechanics is valid. Thus the classical action function W is an approximation for $-(h/2\pi i)\log\psi$. This fact may be used as the basis for a new proof of the basic proposition that suitable wave packet solutions of Eq. (I, 38) obey the laws of classical mechanics and this gives a justification of Eq. (I, 38) independent of that given in Section 1, Part I.[4a]

Turning now to the problem of the variable external field we see that a rational method of setting up the new wave equation is to construct the corresponding Hamiltonian function in Cartesian coordinates, convert it into an operator as before and finally use the operator as in Eq. (1) above.

Let $A_x^{(j)}$, $A_y^{(j)}$, $A_z^{(j)}$, denote the components of the vector potential evaluated at the point x_j, y_j, z_j where the jth electron is located. Let the algebraic value of the electronic charge be e. Then the classical Hamiltonian for a system of τ electrons is[5]

[4a] For a direct and comprehensive proof of this proposition see also Sect. 6, 5 and Sect. 7, 1.

[5] Cf. M. Born, "Atommechanik" Sec. 34, p. 239. Many authors, including Born, use e for the absolute value of the electronic charge and so reverse the sign of the terms linear in e and A both in the Hamiltonian and the wave equation.

$$H(q,p)=\sum_{j=1}^{\tau}\frac{1}{2\mu}\left\{\left(p_{x_j}-\frac{e}{c}A_x^{(j)}\right)^2+\left(p_{y_j}-\frac{e}{c}A_y^{(j)}\right)^2+\left(p_{z_j}-\frac{e}{c}A_z^{(j)}\right)^2\right\}+V(x_1,\cdots z_\tau). \quad (5)$$

Making the substitution $p_k\rightarrow-(h/2\pi i)\partial/\partial q_k$ we obtain as the desired wave equation

$$\sum_{j=1}^{\tau}\left\{\left(\frac{\partial}{\partial x_j}+\frac{2\pi ei}{hc}A_x^{(j)}\right)^2+\left(\frac{\partial}{\partial y_j}+\frac{2\pi ei}{hc}A_y^{(j)}\right)^2+\left(\frac{\partial}{\partial z_j}+\frac{2\pi ei}{hc}A_z^{(j)}\right)^2\right\}\psi-\frac{8\pi^2\mu}{h^2}V\psi-\frac{4\pi\mu i}{h}\frac{\partial\psi}{\partial t}=0. \quad (6)$$

If we make the substitution $\psi=e^{-2\pi iW'/h}$, multiply by h^2, and allow h to approach zero, Eq. (6) yields the Hamilton-Jacobi equation for W' as in the case of a steady electrostatic field. From this fact we infer that Eq. (6) is in satisfactory agreement with classical mechanics in the limit when h is zero.

Expansion of the squared operators throws the differential equation into the form

$$\sum_{j=1}^{\tau}\left[\nabla_j^2\psi+\frac{4\pi ei}{hc}\left(A_x^{(j)}\frac{\partial\psi}{\partial x_j}+A_y^{(j)}\frac{\partial\psi}{\partial y_j}+A_z^{(j)}\frac{\partial\psi}{\partial z_j}\right)+\frac{2\pi ei}{hc}\left(\frac{\partial A_x^{(j)}}{\partial x_j}+\frac{\partial A_y^{(j)}}{\partial y_j}+\frac{\partial A_z^{(j)}}{\partial z_j}\right)\psi\right]-\frac{8\pi^2\mu}{h^2}V'\psi-\frac{4\pi\mu i}{h}\frac{\partial\psi}{\partial t}=0 \quad (7)$$

where [6] $V'=V+(e^2/2\mu c^2)\sum_{j=1}^{\tau}|A^{(j)}|^2$.

Referring to Eq. (I, 102) and to the discussion of Eq. (I, 149) on p. 207 Part I, we observe that to retain our fundamental physical interpretation of $\psi\,\psi^*dv$ it is necessary that

$$\mu\frac{\partial}{\partial t}(\psi\psi^*)=-\operatorname{div}F, \quad (8)$$

where F is a function which vanishes at infinity with ψ and which plays the rôle of mass current density. Using the modified wave equation (7) for a single particle (three dimensional case) we find that Eq. (8) is satisfied with[7]

$$F=\frac{h}{4\pi i}[\psi\ \operatorname{grad}\psi^*-\psi^*\ \operatorname{grad}\psi]-\frac{e}{c}A\psi\psi^*. \quad (9)$$

[6] The correction term $(e^2/2\mu c^2)\sum_{j=1}^{\tau}|A^{(j)}|^2$ is usually negligible.

[7] In Eq. (I, 102) which defines the vector current density, the sign of the right hand member should be reversed.

This is in agreement with the fact that classically the momentum $\boldsymbol{p}$ of a particle in a field with vector potential $\boldsymbol{A}$ is not μv, but $\mu v + eA/c$. Hence the operator associated with μv should be

$$-\frac{h}{2\pi i}\text{grad}-\frac{e}{c}\boldsymbol{A}.$$

From Eqs. (8) and (I, 78) we may deduce the validity of Eq. (I, 171) for the modified operator $H(q, -\partial/\partial q)$ and hence the Hermitian character of the modified matrix $H(n, m)$. Moreover, the whole discussion carries over without difficulty to the case of a system of τ particles.[8]

3. The Perturbation of an Atomic System by a Radiation Field: Absorption

Consider next the application of the wave equation (7) to the problem of the perturbation of an atomic system by an external radiation field.[9] The general solution ψ of (7) can be described in terms of the unperturbed characteristic functions by the aid of an expansion of the form

$$\psi = \sum_{n,m} c_{nm} u_{nm}(x) e^{2\pi i E_{nm} t/h} + \int \sum_{m} c_m(E) u_m(x, E) e^{2\pi i E t/h} dE. \qquad (10)$$

(Here wave functions carry a double subscript to indicate degeneracy. In practice a more complicated notation may be necessary.) The coefficients c_{nm} and c_m(E) become functions of the time during the period of application of the radiation field but settle down to a new set of permanent values different from the initial set when the field is removed. Our first problem is to correlate this change in the c's with the experimental absorption of light and to use the theory to derive theoretical values of Einstein's "transition probabilities." Due to the variation of the c's with time, we shall also find that the wave function executes oscillations which contain the frequencies characteristic of the radiation field, the natural frequencies of the atom, and combinations of the two. Our second problem is to use these three types of frequency to account for dispersion, normal scattering, emission, and Raman scattering.

As previously explained (Sect. 4, 3), Born interpets the ψ functions as descriptions of the behaviour of ensembles of atomic systems subjected to identical external influences. He assumes that the energy of an arbitrary system taken from the ensemble must in every case have an unique value identical with one of the familiar energy levels which play such an important part in the Bohr theory. Finally in the case of a non-degenerate system with a

[8] A wave equation for a single electron in a variable electromagnetic field was first derived by Gordon, Zeits. f. Physik **40**, 117 (1926), and independently by Klein, Zeits. f. Physik **41**, 407 (1927) in connection with their relativistic generalizations of the Schrödinger theory. Although the Gordon-Klein theory does not satisfy the requirements of Eq. (8) and hence involves a new definition of number density to take the place of $\psi\psi^*$, Eq. (6) in three dimensions can be derived from it by a suitable limiting process.

[9] The reader is referred to the article by K. L. Wolf and K. F. Herzfeld, "Absorption and Dispersion," Handbuch der Physik, XX, J. Springer, Berlin, 1928, for a review of this subject from both the experimental and theoretical standpoints.

discrete spectrum only, he identifies the probability of the energy E_n with the corresponding product $c_n c_n^*$.[10] This interpretation is based on the theorem that in the absence of a continuous spectrum

$$\sum_n c_n c_n^* = 1.$$

(Cf. Part I, Eqs. (81, 137, 172)). The generalization required to take into account a continuous spectrum will be evident from Eq. (I, 81).

These assumptions are in comformity with the discussion of Part I, Sect. 1,8 and seem to the authors quite reasonable.

A complete analysis of Born's assumptions cannot be given here, but the following suggestions may be helpful to the reader:

(a) Born supposes that an exact measurement of E for a single system is conceivable and would give a definite result. In support of this assumption we observe that photoelectric measurements with different wave-lengths indicate that we may determine the energies of photons by spectrum analysis, attributing to each photon a definite energy corresponding to the place at which it impinges on the photographic plate in an ideal spectroscopic experiment. In the same way a magnetic deflection experiment yields an energy spectrum for free electrons with a definite energy for each particle.

(b) In Part I, Sect. 1, 8 we have shown that if we perform a Fourier analysis of the wave function for a free particle into terms harmonic in the independent variables x, y, z the amplitudes give the probabilities of the conjugate dynamical variables p_x, p_y, p_z. Furthermore, from the wave standpoint a spectrometer is a mechanical device for analyzing a beam of radiation into a sum of terms which are harmonic in the time. The probability of each elementary energy range for an arbitrary photon is then fixed by the intensity of the corresponding component in the Fourier analysis with respect to time. Hence in the case of a system whose Hamiltonian function does not contain the time explicitly and in which an expansion of the type

$$\psi = \sum_n c_n u_n(x)\, e^{2\pi i E_n t/h} \qquad (a)$$

is possible, a plausible extrapolation from the above special cases suggests that we identify $c_n c_n^* u_n(x) u_n^*(x) d\tau$ with the probability that the system has the energy E_n and also a configuration in the element $d\tau$ of configuration space. If the functions u_n are normalized, integration over all of configuration space gives $c_n c_n^*$ as the probability of the energy E_n independent of the configuration. This is perhaps sufficient justification for Born's assumption in the case under consideration.

(c) On the other hand, in the case of a system whose Hamiltonian does contain the time explicitly so that an expansion of the type (a) is valid only if the c_n's vary with the time, it is evident that Eq. (a) does not describe an exact Fourier analysis of the wave function. Moreover, energy and time are conjugate dynamical variables subject to the Heisenberg uncertainty principle, $\delta E \delta t \sim h$, and so the measurement of either the energy of a system or its energy distribution function must take an infinitely long time. Thus a rigorous non-arbitrary meaning cannot be attached to the phrase "instantaneous probability of the energy E_n." In practice, however, we do not expose photographic plates for an infinitely long time in spectroscopic measurements, and in the case of a fluctuating light source we understand by the frequency distribution at a given time the distribution obtained on a photographic plate exposed for a short time before and after the instant in question. This distribution depends on the duration of exposure as well as the time at which it is started, but is sufficiently definite for practical purposes. This practical procedure of exposing a plate for a finite time, say $t_0 \leqq t \leqq t_1$ is equivalent to the mathematical process of replacing a given function $f(t)$ by a second function $\phi(t)$ defined by

[10] This statement assumes that the expansion is made in terms of normalized wave functions.

$$\phi(t) = f(t), \text{ if } t_0 \leqq t \leqq t_1;\ \phi(t) = 0, \text{ if } \begin{matrix} t < t_0 \\ t > t_1, \end{matrix}$$

and performing a Fourier analysis on $\phi(t)$. For every pair of values of $t_1 - t_0$ and $\bar{t} = (t_1 + t_0)/2$ we get a definite distribution function which has a certain claim toward the title "distribution function for $t = \bar{t}$." Applying this idea to the case of an expansion of the type (*a*) where the c's vary slowly with time, we see that the instantaneous energy distribution function will necessarily be continuous but will have sharp maxima and minima near the energies E_n of the unperturbed system and will, in fact, be an approximation to the discontinuous spectrum postulated by Born's hypothesis. It can also be proved that if the c's are suddenly stabilized at any given instant by removing the perturbations which cause the c's to vary and if we determine by Fourier analysis the energy distribution function for all future time, we shall get exactly Born's result. In other words, Born's identification of $c_n c_n^*$ at the instant t with the instantaneous probability of the energy E_n is mathematically sound if we define the latter quantity as the probability of E_n, for all future time, *which would be created by removal of the perturbation at the instant in question.*

(d) It should be emphasized that the adoption of Born's statistical interpretation means that the wave functions have meaning only in connection with ensembles of atoms. It seems clear that if a wave function did exist for an individual atom, it would still be impossible to measure it, although in the case of an ensemble of similar atoms the observation of $\psi\psi^*$ is at least conceivable. If we adopt Bridgman's operational point of view we must then refuse to attribute objective reality to the wave functions except in connection with ensembles of identical systems. This view becomes even more inescapable when one considers the relation between the observed orbits of large scale bodies or of electrons and the associated wave packets. The wave packets must in all cases increase steadily in volume with time just as a blob of radiation emerging from a small aperture increases in volume with time. If, however, we observe that at any instant a particle is in the neighborhood of some particular point this observation modifies our expectation for the future and makes it necessary to discard our original wave-packet and substitute, as a basis for prediction, a new one based on our observation. This substitution of a new wave function for the old makes sense only if we adhere strictly to the view that wave functions are simply tools for the description of the statistical behavior of ensembles of identical mechanical systems.

The reader is referred to a recent paper of Lewis and Mayer, Proc. Nat. Acad. Sci. **15**, 127 (1929) for other very interesting considerations in this line.

On this basis the change in the value of $c_n c_n^*$ produced by any given perturbation measures the fraction of the whole number of systems in the ensemble which pass into the corresponding energy level as a result of the perturbation. We proceed to outline a scheme for computing this change in the case of a radiation field.

For simplicity let us assume that the radiation consists of a system of plane waves moving in the direction of the x-axis and polarized with the electric vector parallel to the y-axis. In conformity with the perturbation theory of Section 4, 3 we assume a finite wave train described for convenience by a complex Fourier integral. Let the vector potential and electric force be

$$\begin{aligned} A_y &= -\alpha f(t - x/c) = +\alpha c \int_{-\infty}^{+\infty} \frac{\phi_y(\nu)}{2\pi i \nu} e^{2\pi i \nu (t - x/c)} d\nu, \\ E_y &= -\frac{1}{c}\frac{\partial A_y}{\partial t} = -\alpha \int_{-\infty}^{+\infty} \phi_y(\nu) e^{2\pi i \nu (t - x/c)} d\nu. \end{aligned} \tag{11}$$

Here α is a real parameter which determines the amplitude of the wave system, while the complex function $\phi_\nu(\nu)$ fixes the wave form and may be subjected to the condition

$$\int_{-\infty}^{+\infty} \phi_\nu \phi_\nu^* d\nu = 1. \tag{12}$$

In the case of a radiation field the divergence of the vector potential vanishes. Using this fact and multiplying Eq. (7) by $-h^2/8\pi^2\mu$ we reduce it to the form (neglecting the term $(e^2/2\mu c^2)\sum_j |A^{(j)}|^2$),

$$\left[H_0 + \frac{h}{2\pi i}\frac{e}{\mu c}\sum_{j=1}^{\tau}\left(A_x^{(j)}\frac{\partial}{\partial x_j} + A_y^{(j)}\frac{\partial}{\partial y_j} + A_z^{(j)}\frac{\partial}{\partial z_j}\right)\right]\psi = \frac{h}{2\pi i}\frac{\partial \psi}{\partial t}, \tag{13}$$

where H_0 is the unperturbed Hamiltonian operator. Introducing the first of Eqs. (11) and identifying the parameter λ of Section 4, 3 with α we obtain the following expression for the perturbing operator $F(t)$:

$$F = -\frac{h}{2\pi i}\frac{e}{\mu c}\left[\sum_{j=1}^{\tau} f(t - x_j/c)\frac{\partial}{\partial y_j}\right]. \tag{14}$$

The first step in carrying through the perturbation scheme of Section 4, 3 is to evaluate the matrix elements

$$F_{mn}(t) = \int u_n^* F u_m d\nu.$$

If the wave-lengths of the radiation field are all large compared with the atomic dimensions we may regard A as constant over the atom and may replace $f(t-x_j/c)$ by the value of $f(t-x/c)$ at the nucleus of the atom in computing $F_{mn}(t)$. If the origin is at the nucleus

$$F_{mn}(t) = \frac{e}{\mu c} f(t)\int_\infty u_n^* \sum_{j=1}^{\tau}\left(-\frac{h}{2\pi i}\frac{\partial}{\partial y_j}\right) u_m d\nu. \tag{15}$$

$\sum_j(-h/2\pi i)\partial/\partial y_j$ will be recognized as the operator associated with the y component of the linear momentum (Cf. Section 3, 2 pp. 192-3). The integral in the right hand member is therefore the matrix element of $p_y(m,n)$ of the matrix defined by the expansion

$$p_y u_m = \sum_n p_y(m,n) u_n. \tag{16}$$

$p_y(m,n)$ in turn is the amplitude of the corresponding element $p_y(m,n/t)$ of the matrix function of t associated with p_y in the theory of Heisenberg, Born, and Jordan. The latter matrix is defined by an expansion similar to (16) with u_m and u_n replaced by ψ_m and ψ_n respectively. Thus

$$p_y(m,n)e^{(2\pi i/h)(E_m - E_n)t} = p_y(m,n/t) = \int_\infty \psi_n^* \sum_{j=1}^{\tau}\left(-\frac{h}{2\pi i}\frac{\partial}{\partial y_j}\right)\psi_m d\nu. \tag{17}$$

Classically the linear momentum is μ/e times the time derivative of the electric moment and a similar theorem holds for the corresponding matrices. Let P_y denote the y component of the electric moment, i.e., $\sum_j e_j y_j$ and let $\boldsymbol{P}_y(m,n/t)$ and $P_y(m,n)$ denote elements of the matrices of P_y based on the functions ψ and u respectively. Then

$$(e/\mu)p_y(m,n/t)=\dot{P}_y(m,n/t)=(2\pi i/h)(E_m-E_n)P_y(m,n/t) \tag{18a}$$

and consequently

$$(e/\mu)p_y(m,n)=(2\pi i/h)(E_m-E_n)P_y(m,n). \tag{18b}$$

With the aid of Eq. (18b) the expression for $F_{mn}(t)$ can be thrown into the form

$$F_{mn}(t)=\frac{2\pi i}{hc}(E_m-E_n)P_y(m,n)f(t). \tag{19}$$

The first order perturbations of the wave functions are now determined from Eq. (I, 179) which, with p set equal to unity, should read[11]

$$f^{(1)}_{mn}(T)=(2\pi i/h)\int_0^T F_{mn}(t')e^{2\pi i\nu_{mn}t'}dt' \tag{20}$$

where ν_{mn} is defined by

$$h\nu_{mn}=E_m-E_n=-h\nu_{nm}.$$

By combining Eqs. (19) and (20) and introducing the complex Fourier coefficient ϕ_y of Eq. (11) we obtain[12]

$$f^{(1)}_{mn}(T)=-\left(\frac{2\pi i}{h}\right)P_y(mn)\phi_y(\nu_{nm}), \tag{21}$$

and finally in first approximation

$$c_n=c_n{}^0-\alpha\sum_m c_m{}^0\left(\frac{2\pi i}{h}\right)P_y(mn)\phi_y(\nu_{nm})+\cdots. \tag{22}$$

[11] Unfortunately the factor $e^{2\pi i\nu_{kn}t'}$ was accidentally omitted in preparing the manuscript of Part I.

[12] To elucidate this result we find first by substituting from Eqs. (11) and (19) for $f(t)$ and $F_{mn}(t)$ respectively into Eq. (20),

$$f^{(1)}_{mn}(T)=-\left(\frac{4\pi^2}{hc}\right)\nu_{mn}P_y(mn)\int_0^T f(t')e^{2\pi i\nu_{mn}t'}dt'.$$

But from the definition of the Fourier Integral and Eq. (11),

$$\frac{c\phi_y(\nu_{nm})}{2\pi i\nu_{nm}}=\int_{-\infty}^{+\infty} f(t')e^{2\pi i\nu_{mn}t'}dt'.$$

The replacement of the integral with respect to t' over the region 0 to T by an integral from $-\infty$ to $+\infty$ is valid because we have assumed explicitly $f(t')\equiv 0$ unless $0\leqq t'\leqq T$.

The same result is obtained somewhat more directly, but less rigorously, if one treats the product $E_y P_y$ as a perturbing potential function depending explicitly on the time.

In case the radition field is not plane polarized, Eq. (22) is to be replaced by

$$c_n = c_n^0 - \alpha \sum_m c_m^0 \left(\frac{2\pi i}{h}\right)\left(P(mn)\phi(\nu_{nm})\right) + \cdots \tag{23}$$

where $P(mn)$ and $\phi(\nu_{nm})$ are vectors derived from the vector electric moment and vector electric force respectively.[13]

It is convenient at this stage of the argument to introduce the quantity Φ_{mn} defined by the equation

$$\Phi_{mn} = f_{mn}(T) f_{mn}^*(T). \tag{24}$$

In case all the atoms are initially in the same state m Eq. (I, 180) shows that Φ_{mn} measures the fraction of the atoms transferred from the state m to the state n by the perturbing field. Hence we may call Φ_{mn} a transition probability for jumps from the state m to the state n. Some analysis is required, however, before we can make the connection between Φ_{mn} and the corresponding transition probability as defined by Einstein. The derivation of the blackbody radiation formula by the latter author is based on the assumption that the rate of increase of the number of atoms, N_n, in the state n due to absorption in a natural radiation field is given by an expression of the form

$$\frac{dN_n}{dt} = \sum_m (N_m B_{m\to n} - N_n B_{n\to m}) u(\nu_{nm}) \tag{25}$$

where $u(\nu_{nm})$ is the monochromatic energy density of the field, and the B's are the probabilities of jumps from the state m to the state n and from n to m respectively, per atom per unit time when $u(\nu_{nm})$ is unity. It is evident that if the present theory is to harmonize with Einstein's it must be possible to derive Eq. (25) from Eq. (23). Since Einstein's theory ignores fluctuations from thermodynamic equilibrium we can identify N_n in Eq. (25) with the product of the number of atoms per unit volume, N, by the mean value of $c_n c_n^*$ obtained by averaging over all the atoms and also over an interval of time long enough to eliminate all high frequency terms.[14] $(1/N)dN_n/dt$ then becomes identical with the average "secular" rate of increase of $\overline{c_n c_n^*}$, and we have to prove that

$$\frac{d}{dt}\overline{c_n c_n^*} = \frac{d}{dt}\overline{\overline{c_n c_n^*}} = \sum_m \left[\overline{\overline{c_m c_m}}^* B_{m\to n} - \overline{\overline{c_n c_n^*}} B_{n\to m}\right] u(\nu_{nm}). \tag{26}$$

[13] The large round brackets indicate a scalar product.

[14] We here treat each individual atom in any small volume element of a gas by the wave mechanics, neglecting all interactions. Each individual atom in the element then has its own ψ function correlated with an imaginary ensemble of similar independent atoms. Although the different atoms in the volume element are physically indistinguishable, a confusion of the group actually present in the element with the ensemble which defines the ψ function of any one of them is undesirable. Cf. the discussion on p. 15.

It is also permissible to restrict the proof to the case where the perturbing radiation field is "natural." This restriction is interpreted to mean that the field consists of many superposed short wave trains whose phases are distributed at random.

By Eq. (I, 180) there are terms of the form $c_l^0 c_n^{0*} f_{ln}(T) f_{ln}^*(T)$ in the expression for $c_n(T)c_n^*(T)$, but these average to zero for the special conditions under consideration and it is possible to prove[15] that Eq. (26) holds in first approximation with

$$B_{m\to n}u(\nu_{nm}) = \frac{\Phi_{mn}}{T} = \frac{8\pi^3}{3h^2} \left| \boldsymbol{P}(mn) \right|^2 u(\nu_{nm}), \tag{27}$$

or

$$B_{m\to n} = B_{n\to m} = \frac{8\pi^3}{3h^2} \left| \boldsymbol{P}(mn) \right|^2. \tag{28}$$

In virtue of Eq. (18) the above theoretical expression for the Einstein transition probability $B_{m\to n}$ can be expressed in the alternative form

$$B_{m\to n} = \frac{2\pi e^2}{3\mu^2} \frac{\left| \boldsymbol{p}(m,n) \right|^2}{(E_m - E_n)^2} \tag{29}$$

provided that E_n is not equal to E_m.

The theory is applicable to degenerate systems if we interpret $B_{m\to n}$ as the transition probability from one elementary state to another and not from one energy level to another. For such systems it is desirable to introduce a more complete notation in which each of the symbols m and n is replaced by a pair of symbols one of which designates the energy level while the other indicates the particular sub-state under consideration. $B_{m\to n}$ then goes over into $B(m, l\to n, k)$ with m and n designating the energy levels. Eq. (25) becomes

$$\frac{d}{dt}N_{nk} = \sum_{m,l} \left[N_{ml}B(m,l\to n,k) - N_{nk}B(n,k\to m,l) \right] u(\nu_{nm}). \tag{30}$$

Summing both sides over all values of k (i.e., over all sub-states of the nth energy level) we obtain

$$\frac{dN_n}{dt} \equiv \frac{d}{dt}\left(\sum_k N_{nk} \right) = \sum_m \left[N_m \overline{B}_{m\to n} - N_n \overline{B}_{n\to m} \right] u(\nu_{nm}) \tag{31}$$

with

$$\left.\begin{aligned} \overline{B}_{m\to n} &= \sum_{l,k} \frac{N_{ml}}{N_m} B(m,l\to n,k) = \frac{1}{g_m} \sum_{l,k} B(m,l\to n,k) \\ \overline{B}_{n\to m} &= \sum_{l,k} \frac{N_{nk}}{N_n} B(n,k\to m,l) = \frac{1}{g_n} \sum_{l,k} B(n,k\to m,l). \end{aligned}\right\} \tag{32}$$

[15] M. Born, Zeits. f. Physik **40**, 167 (1926); J. C. Slater, Proc. Nat. Acad. Sci. **13**, 7 (1927).

Here g_m denotes the statistical weight of the energy level E_m, and is equal to the number of degenerate levels associated with symbol m, thus being equal to N_m/N_{ml}. Eq. (31) agrees with the Einstein form for degenerate systems, and Eqs. (32) are in conformity with the familiar relation

$$\overline{B}_{n\to m}=(g_m/g_n)\overline{B}_{m\to n}.$$

Eq. (19) shows that a radiation field has no tendency to produce transitions from one member of a degenerate group of states to another member of the same group.

Eq. (27) shows that to a first approximation the intensity of any absorption line is proportional to the square of the absolute value of the corresponding matrix element of the dipole moment of the atomic system. To get the absorption (and hence emisson) due to the quadrupole moment which is important in connection with the appearance of "forbidden" lines in nebular spectra[16] one must take into account the variation in the vector potential $\boldsymbol{A}$ over the volume of the atom, which was neglected in deriving Eq. (15).

Evaluation of the matrix elements $\boldsymbol{P}(mn)$ or $\boldsymbol{p}(mn)$ yields the familiar selection principles for the magnetic and azimuthal quantum numbers of atomic systems.[17] In the case of the two-body problem, for example, using the wave functions of Sect. 2, 6 we obtain

$$P_z(n',l',m';n'',l'',m'')=e\left[\int_0^\infty R_{n',l'}R_{n'',l''}r^3dr\right]\left[\int_0^\pi \Theta_{l',m'}\Theta_{l'',m''}\sin\theta\cos\theta d\theta\right]$$

$$\times\left[\int_0^{2\pi}\Phi_{m'}\Phi^*_{m''}d\phi\right]=e\left[\int_0^\infty R_{n'l'}R_{n''l''}r^3dr\right]\delta_{l',l''\pm1}\delta_{m',m''}, \tag{33}$$

with similar expressions for the matrix components of P_y and P_x.

The theory can be extended to include the continuous part of the absorption spectrum as well as the discontinuous part. To this end the perturbation theory of Section 4, 3 must be completed by the inclusion of the continuous characteristic functions in the expansion of the perturbed function [Eq. (68)]. For details the reader is referred to the original papers of Wentzel and others.[18]

4. Emission, Dispersion, and Scattering

Einstein's derivation of the black body radiation formula is based on the assumption that the mean rate of emission per unit volume of the frequency ν_{mn} associated with the quantum jump $m\to n$ in a gas in thermal equilibrium is given by an expression of the form

$$S(\nu_{mn})=N_mA_{m\to n}h\nu_{mn} \tag{34}$$

where N_m is the average number of atoms or molecules per unit volume in the state m, and $A_{m\to n}$ is a constant dependent only on the nature of the gas.

[16] Cf. James H. Bartlett, Phys. Rev. **34**, 1247 (1929).

[17] For details see L. Brillouin, Journ. de Physique **8**, 74 (1927).

[18] G. Wentzel, Zeits. f. Physik **40**, 574 (1926). J. R. Oppenheimer, Zeits. f. Physik **41**, 268 (1927); Phys. Rev. **31**, 66 (1928). P. A. M. Dirac, Proc. Roy. Soc. **A113**, 621 (1927).

The coefficients $A_{m\to n}$ are called "transition probabilities for spontaneous emission" and are related to the corresponding absorption probabilities by the equation

$$A_{m\to n}=(8\pi h\nu_{mn}^3/c^3)B_{m\to n}. \tag{35}$$

Combining Einstein's well-grounded and widely accepted theory with the wave-mechanics formula for $B_{m\to n}$ developed in the preceding article we arrive at an indirect method for computing $S(\nu_{mn})$, and hence the intensities of emission lines in gaseous spectra, applicable to all cases in which Eq. (34) is valid and values of N_m are obtainable. In the case of purely thermal radiation where the departure from thermodynamic equilibrium is small, the "population numbers" N_m can be estimated from the Maxwell-Boltzman law. In other cases these numbers are uncertain, but it is reasonable to assume in the case of an electric discharge that the populations of states whose energies are nearly the same are proportional to the statistical weights, as they are for thermal equilibrium. This assumption seems to be in accordance with the experimental facts. At the worst we can test the theory under conditions in which the population numbers are unknown by measuring and computing the relative intensities of the different spectrum lines originating in a common upper energy level.

A more direct method of computing emission intensities is highly desirable, however, and in the last analysis necessary. Eq. (34) asserts in effect the incoherence of the radiation fields of neighboring atoms and is therefore inapplicable when the electromagnetic waves sent out by neighboring atoms have definite phase relationships. Consider, for example, the metallic reflection of resonance radiation observed by Wood at the surface of a resonance bulb of mercury vapor.[19] The regularity of the reflection and the polarization of the reflected light show clearly that in this case the emission of the surface atoms is coherent so that amplitudes add, rather than intensities. By lowering the pressure, however, there is a continuous transition between this regular reflection and the emission of diffuse incoherent fluorescence light by the irradiated vapor. Similarly in the case of ordinary dispersion we have to do with elementary wavelets emitted by adjacent atoms with definitely related phases although the frequency emitted is a frequency of forced vibration which does not happen to coincide with any of the transition frequencies of the atoms.

As a first step toward the development of a theory of emission applicable to coherent as well as incoherent conditions we observe that since the ψ functions give a statistical description of the atomic motions while the electric and magnetic vectors give a corresponding statistical description of the emitted streams of photons, it is to be expected *a priori* that $\boldsymbol{E}$ and $\boldsymbol{H}$ for the emitted radiation shall be determined by the ψ functions just as in the classical theory $\boldsymbol{E}$ and $\boldsymbol{H}$ were determined by the motion of the electrical charges. Schrödinger's original suggestion was that the charge density ρ to be used in

[19] R. W. Wood, Phil. Mag. (6) **18**, 187 (1909); (6) **23**, 689 (1912); R. W Wood and M. Kimura, Phil. Mag. (6) **32**, 329 (1916).

Maxwell's equations should be identified, in the case of a single electron, with $e\psi\psi^*$. From our present point of view this means that ρ is to be identified with the statistical average charge per unit volume at the point in question. If the vector current density for a single electron is identified with

$$\frac{eh}{4\pi\mu i}[\psi \operatorname{grad}\psi^* - \psi^* \operatorname{grad}\psi] \tag{36}$$

[Cf. Part I, Section 3, 2] the law of continuity is satisfied, as well as the law of the conservation of electricity. In the case of a system of many electrons $\rho(x, y, z)$ and the current density are to be obtained by summing up the contributions of all the individual charges. The formulas are cumbersome but need not be introduced explicitly here since in practice the radiation field of a microscopic emitter is sufficiently determined by its dipole moment. In conformity with the above definition of ρ one may compute the statistical mean value of the electric moment $\boldsymbol{P}$ from the formula

$$\begin{aligned}\overline{\boldsymbol{P}}(t) &= \int\int\int_{\infty} \boldsymbol{r}\rho(x,y,z)\,dx\,dy\,dz \\ &= \int_{\text{all coordinate space}} \sum_{j=1}^{n} e\boldsymbol{r}_j \psi\psi^*\,d\tau\end{aligned} \tag{37}$$

where $\boldsymbol{r}$ is the radius vector from the origin to the point (x, y, z).

Expanding ψ in terms of the characteristic functions and neglecting the continuous spectrum as usual, we obtain

$$\overline{\boldsymbol{P}}(t) = \sum_n \sum_m c_n^* c_m \boldsymbol{P}(m, n/t) = \sum_n \sum_m c_n^* c_m \boldsymbol{P}(m,n) e^{2\pi i \nu_{mn} t}, \tag{38}$$

where $\boldsymbol{P}(m, n/t)$ and $\boldsymbol{P}(m,n)$ are elements of the vector matrices of $\boldsymbol{P}$ based on the functions $\psi(x, t)$ and $u(x)$ respectively. (Cf. p. 8.) The matrix $\boldsymbol{P}(m, n/t)$ is Hermitian so that each term in the above sum is the complex conjugate of the term obtained by interchanging the values of m and n. Hence $\overline{\boldsymbol{P}}(t)$ is real, as it should be. Eq. (38) shows that the frequencies of the electric moment as defined are identical with those given by the Bohr frequency condition and hence gives initial assurance that we are on the right track.

To determine the mean rate of emission of energy per atom in an ensemble with wave function ψ we insert $\overline{\boldsymbol{P}}(t)$ for the electric moment in the classical electrodynamic formula

$$\bar{I} = \frac{2}{3c^3}\left|\frac{d^2\overline{\boldsymbol{P}}}{dt^2}\right|^2 = \frac{2}{3c^3}\overline{[(\ddot{\overline{P}}_x)^2 + (\ddot{\overline{P}}_y)^2 + (\ddot{\overline{P}}_z)^2]} \tag{39}$$

Using the series of Eq. (38) and reducing, we obtain

$$\bar{I} = \frac{4}{3c^3} \sum_{\nu mn} \sum_{>0} |c_n|^2 |c_m|^2 (2\pi\nu_{mn})^4 |\boldsymbol{P}(mn)|^2.$$

Defining $A_{m\to n}$ by Eqs. (35) and (28), this becomes

$$\bar{I} = \sum_{\nu_{mn}>0} \sum | c_n |^2 | c_m |^2 A_{m\to n} h\nu_{mn}. \tag{40}$$

Thus we find the emitted stream of energy made up of a sum of terms each having one of the frequencies of the actual spectrum and proportional to the corresponding transition probability $A_{m\to n}$. The theory is in harmony with the classical electromagnetic theory in the limiting case of a macroscopic system whose configuration can be sharply defined by a suitable wave packet since in this case $\bar{P}(t)$ reduces to the classical electric moment.

When, however, we apply our equation to the emission of radiation by a gas we strike a snag. Strictly speaking, we should treat the entire sample of gas as a single electro-mechanical system and compute the emission from a single function describing an ensemble of such systems. As this method of attack is too difficult, we content ourselves with a short cut suggested by classical considerations. Let us assume that the radiation from any element of volume can be obtained by compounding the electric moments of a set of independent atoms each of which is a member of a separate ensemble with a separate ψ function. The electric moment of the volume element $dxdydz$ is then the sum of the values of $\bar{P}(t)$ for all atoms in the element. If the phases of the $\bar{P}$'s are haphazard, the intensities of emission add and the mean rate of emission of frequency ν_{mn} is[20]

$$\sum_{\text{All atoms}} | c_n |^2 | c_m |^2 A_{m\to n} h\nu_{mn}.$$

Let N denote the number of atoms per unit volume. Then the rate of emission of the frequency ν_{mn} per unit volume is

$$S(\nu_{mn}) = N \overline{| c_m |^2 | c_n |^2} A_{m\to n} h\nu_{mn}, \tag{41}$$

where the bar denotes an average over the atoms in the element. Clearly $N \overline{c_m{}^2}$ may be identified with the number of atoms per unit volume in the state m. Were it not for the factor $|c_n|^2$ in the right hand member, Eq. (41) would be in satisfactory agreement with Eq. (34). This factor, however, makes the emission of energy of any frequency depend on the number of atoms in the corresponding lower energy level as well as on the number in the upper energy level, and is in violent conflict with Einstein's theory and with experience.

The difficulty here encountered is fundamental and apparently can be met only by a modification of our conception of the electromagnetic field. The necessity for such modification was foreshadowed in early papers on the matrix mechanics by Born and Jordan, and by Born, Heisenberg, and Jordan[21] who assumed from the beginning that the kinematics of the electric and

[20] In order to take into account the effect of collisions between atoms, the perturbations by the ever present radiation field, or the damping effect of radiation resistance, one must regard the c_m's and c_n's as functions of the time. In that case it would be necessary to average the product with respect to time as well as over all the atoms in the volume element.

[21] M. Born and P. Jordan, Zeits. f. Physik **34**, 858 (1925); M. Born, W. Heisenberg, and P. Jordan, Zeits. f. Physik **35**, 557 (1926).

magnetic vectors must be the same as that of the coordinates and momenta of matter particles. To be specific, they assumed the validity of Maxwell's equations for the electromagnetic field but re-interpreted the dependent variables $\boldsymbol{E}$ and $\boldsymbol{H}$ in these equations as matrix functions of the coordinates instead of ordinary vector point functions. From the point of view of this review, this means that they regard $\boldsymbol{E}$ and $\boldsymbol{H}$ as quantities which are only partially determinate like the p's and q's in the wave mechanics of a system of material particles.

The plausibility of this procedure is perhaps most evident in the case of a radiation field in an evacuated rectangular box with perfectly reflecting walls. It is well known[22] that in the case of such a box the classical electromagnetic theory permits the resolution of the most general radiation field into the sum of an infinite series of discrete simple harmonic normal modes of vibration. The instantaneous amplitude of one of these vibrations may be regarded as a normal coordinate of the system and the equations of motion may be thrown into Lagrangian or Hamiltonian form. If $\boldsymbol{q}_n$ is any normal coordinate and $\boldsymbol{p}_n$ is the conjugate momentum, the expressions for the electric force $\boldsymbol{E}$ and vector potential $\boldsymbol{A}$ are of the form

$$\boldsymbol{E} = \sum_n \boldsymbol{q}_n f_n(x, y, z) = \sum_n \boldsymbol{E}_n ;$$

$$\boldsymbol{A} = \sum_n \boldsymbol{p}_n a_n f_n(x, y, z) = \sum_n \boldsymbol{A}_n .$$

The analogy between the equations of motion of the normal coordinates of this electromagnetic field and the equations of motion of the coordinates of a system of particles suggests at once that we treat the former like the latter and set up a wave equation based on the Hamiltonian function similar to that used in the treatment of a system of particles. The Heisenberg uncertainty principle then applies to the conjugate variables $\boldsymbol{p}_n$ and $\boldsymbol{q}_n$ and hence in a slightly modified form to the corresponding values of $\boldsymbol{E}_n$ and $\boldsymbol{A}_n$.

Lorentz[22] has further shown that classically the same normal coordinates for the radiation field can be used to describe an interaction between matter and radiation in a box with perfectly reflecting walls, and Dirac[23] has been notably successful in dealing with the problem of the emission and absorption of radiation on the basis of quantum mechanics applied to matter and electromagnetic field together, using the device of the perfectly reflecting box to reduce the equations of motion to Hamiltonian form. In this way he derives the above formulas for the transition probabilities $A_{m\to n}$ and $B_{m\to n}$, and a theory of dispersion in a single self-consistent theory. The authors will not attempt, however, to review here these most important, but somewhat abstruse, papers of Dirac, primarily because up to this point we have not developed the necessary mathematical machinery.

[22] Cf. H. A. Lorentz in the report of the Solvay Conference of 1911: "La Theorie du Rayonnement et les Quanta," p. 12, Paris, 1912. The discussion by Lorentz is not limited to the case where the box is completely evacuated.

[23] P. A. M. Dirac, Proc. Roy. Soc. **A114**, 243 and 710 (1927).

It will be worth while, however, to examine briefly a less radical procedure recently suggested by Frenkel[24] which is based on the early discussion of the electromagnetic field from the matrix point of view by Born and Jordan.[21] The fundamental idea is that since $\overline{P}(t)$ [Cf. Eqs. (37) and (38)] is the statistical mean value of the electric moment of the emitting system, the electric and magnetic forces computed from it by classical formulas may be expected to be statistical mean values of the actual electric and magnetic forces for individual systems in the ensemble. To determine the corresponding rate of emission we have to compute the statistical mean value of the Poynting vector and then integrate over all directions of emission. If we now use the classical formula

$$\mathfrak{S}=\frac{c}{4\pi}E\times H \tag{42}$$

for the computation of $\overline{\mathfrak{S}}$ from $\overline{E}$ and $\overline{H}$ we obtain the incorrect expression (40) for the mean rate of emission. On the other hand we may proceed as follows. First we use the classical electromagnetic formulas[25]

$$E\left(t+\frac{r}{c}\right)=\frac{1}{c^2r}\left[r\left(r_0\cdot\frac{d^2P}{dt^2}\right)-\frac{d^2P}{dt^2}\right];\quad H\left(t+\frac{r}{c}\right)=\frac{1}{c^2r}\left[r_0\times\frac{d^2P}{dt^2}\right] \tag{43}$$

to compute from each matrix component $P(m, n/t)$ of P corresponding matrix elements of E and H. Although these elements are functions of time we indicate them by simple subscripts as E_{mn} and H_{mn} repectively. From the definitions of P and of matrix elements it follows that

$$\overline{E}=\sum_n\sum_m c_n^*c_mE_{mn};\quad \overline{H}=\sum_n\sum_m c_n^*c_mH_{mn} \tag{44}$$

in analogy with Eq. (38). We now compute the Poynting vector $\overline{\mathfrak{S}}$ from the modified formulas

$$\overline{\mathfrak{S}}=\sum_n\sum_m c_n^*c_m\mathfrak{S}_{mn} \tag{45}$$

and [26]

$$\mathfrak{S}_{mn}=\frac{c}{8\pi}\sum_k(E_{kn}\times H_{mk}-H_{kn}\times E_{mk}). \tag{45a}$$

Forming the time average of the statistical average $\overline{\mathfrak{S}}$, we find that

$$\overline{\overline{\mathfrak{S}}}=\sum_n|c_n|^2\mathfrak{S}_{nn}=\frac{c}{4\pi}\sum_n\sum_k(|c_n|^2+|c_k|^2)(E_{kn}\times H_{nk}). \tag{46}$$

[24] J. Frenkel, "Einführung in die Wellenmechanik," Berlin, 1929, Chap. III, §5. Frenkel's volume reached us after the completion of Part I of this review.

[25] Here r_0 denotes an unit vector along the radius drawn from the atomic center to the point for which E and H are to be calculated.

[26] By definition $(E_{kn}\times H_{mk})^{(x)}=(E_{kn}^{(y)}H_{mk}^{(z)}-E_{kn}^{(z)}H_{mk}^{(y)})$ where $E_{kn}^{(y)}$ is the yth component of the vector matrix E_{kn}, etc. Cf. also footnote 28.

At this point Frenkel proceeds to take the unjustified step of replacing $(|c_n|^2+|c_k|^2)$ by $2|c_n|^2$. With this arbitrary alteration one gets

$$\bar{I}=\sum_{\nu_{nk}>0}\sum|c_n|^2 A_{n\to k}h\nu_{nk} \tag{47}$$

for the mean rate of emission in agreement with Eq. (34).

The above method of computing $\bar{\mathfrak{S}}$ up to Eq. (47) is exactly what we should expect to employ if the vectors $\boldsymbol{E}$, $\boldsymbol{H}$, and $\mathfrak{S}$ for the field of the atom are, like the electric moment $\boldsymbol{P}$ itself, definite functions of the p's and q's of the atom. Then the classical formulas for $\boldsymbol{E}$, $\boldsymbol{H}$, and $\mathfrak{S}$ in terms of the p's and q's determine operators $\boldsymbol{O}_E(x, y, z, t)$, $\boldsymbol{O}_H(x, y, z, t)$, and $\boldsymbol{O}_{\mathfrak{S}}(x, y, z, t)$ which may be used to compute $\bar{\boldsymbol{E}}$, $\bar{\boldsymbol{H}}$, and $\bar{\mathfrak{S}}$ from the ψ function of the emitting ensemble of atoms. Defining the desired mean values by means of the operators as in Part I, Sect. 3 and expanding ψ in terms of the characteristic functions we obtain

$$\begin{aligned}\bar{\boldsymbol{E}}&=\int\psi^*\boldsymbol{O}_E\psi d\tau=\sum_{n,m}c_n^*c_m\boldsymbol{E}_{mn}\\ \bar{\boldsymbol{H}}&=\int\psi^*\boldsymbol{O}_H\psi d\tau=\sum_{n,m}c_n^*c_m\boldsymbol{H}_{mn}\\ \bar{\mathfrak{S}}&=\int\psi^*\boldsymbol{O}_{\mathfrak{S}}\psi d\tau=\sum_{n,m}c_n^*c_m\mathfrak{S}_{mn}.\end{aligned} \tag{48}$$

The matrix elements $\boldsymbol{E}_{mn}$, $\boldsymbol{H}_{mn}$ as defined by Eqs. (48) may readily be proved to be identical with the elements $\boldsymbol{E}_{mn}$, $\boldsymbol{H}_{mn}$ previously defined in terms of $\boldsymbol{P}(m, n/t)$ (at least to the approximation to which Eqs. (43) are valid classically).[27] It remains to show that $\mathfrak{S}_{mn}$ is properly defined by Eq. (45a). The desired proof is a corollary on a general theorem of fundamental importance for the matrix formulation of the quantum mechanics (cf. the next section).

Theorem: Let O_α and O_β denote two operators associated with the dynamical variables α and β. Let α_{mn} be the matrix element defined by the equivalent equations

[27] Classically in the case of a system of electrons of equal mass and charge

$$\frac{d^2\boldsymbol{P}}{dt^2}=\frac{e}{\mu}\sum_j\dot{p}_j=-\frac{e}{\mu}\sum_j\operatorname{grad}_j V$$

where $\operatorname{grad}_j V$ denotes the gradient with respect to the coordinates of the j^{th} particle. We may then use $-(e/\mu)\sum_j\operatorname{grad}_j V$ as an operator for $d^2\boldsymbol{P}/dt^2$ and by combination with Eqs. (43) work out the operators $\boldsymbol{O}_E$ and $\boldsymbol{O}_H$. With the aid of the vector equivalent of Eq. (18a) and a generalization for Eq. (I, 92) it is then possible to show that

$$\frac{d^2\bar{\boldsymbol{P}}}{dt^2}=-\frac{e}{\mu}\overline{\sum_j\operatorname{grad}_j V}$$

and thus establish the equivalence of the two definitions of the matrices for $\boldsymbol{E}_{mn}$ and $\boldsymbol{H}_{mn}$.

$$\alpha_{mn} = \int \psi_n^* O_\alpha \psi_m d\tau \tag{49}$$

$$O_\alpha \psi_m = \sum_n \alpha_{mn} \psi_n . \tag{50}$$

Let β_{mn} be defined in the same way. Finally, let O_γ denote the operator product $O_\alpha O_\beta$. Then

$$\gamma_{mn} = \sum_k \beta_{mk} \alpha_{kn} = \sum_k \alpha_{kn} \beta_{mk} . \tag{51}$$

The right hand member of Eq. (51) is by definition the element (mn) of the matrix product $\{\beta_{mn}\}\ \{\alpha_{mn}\}$ so that the theorem states that the matrix of the product of two operators is the product of the matrices of the operators taken separately, the two products being formed in inverse order.[28]

Proof: It is only necessary to expand $O_\beta \psi_n$ in accordance with Eq. (50) and then apply the operator O_α to each term in the expansion. Then

$$O_\gamma \psi_m = O_\alpha \sum_k \beta_{mk} \psi_k = \sum_k \beta_{mk} \sum_n \alpha_{kn} \psi_n = \sum_n \left(\sum_k \beta_{mk} \alpha_{kn} \right) \psi_n . \tag{52}$$

This proves the theorem subject to the condition that is legitimate to apply the operator O_α term by term to the series $\sum_k \beta_{mk} \psi_k$.

Returning now to the Poynting vector, we observe that in view of the classical formula (42) we may define $O_{\mathfrak{S}}$ by an equation of the form

$$O_{\mathfrak{S}} = \frac{c}{4\pi} O_E \times O_H . \tag{53}$$

Here we may identify O_E and O_H with the vector operators

$$O_E = iO_X + jO_Y + kO_Z$$

$$O_H = iO_\alpha + jO_\beta + kO_\gamma$$

where X, Y, Z, α, β, γ, are the components of $\boldsymbol{E}$ and $\boldsymbol{H}$ respectively. Expanding (53) in terms of the corresponding scalar components we find an

[28] It is well known that the commutative law of ordinary multiplication does not apply to matrix multiplication and obviously it cannot apply to the multiplication of such operators as q and $\partial/\partial q$. For a presentation of the elements of matrix algebra the reader is referred to Bocher's "Higher Algebra" or to any of the current introductions to the matrix mechanics. Cf. also p. 25 ahead.

The inversion of order between the operator products and matrix products is an unhappy peculiarity of the present presentation of the subject, traceable in the first instance to the convention regarding the order of the subscripts in the equations which define the matrices in terms of the operators ((I, 131) and (49) and (50) above). Due to our original unfortunate choice of the sign of the exponent of the time factor of ψ, it was necessary to arrange the subscripts as we have done in order to get the same matrix elements (Hermitian case only) as those used by Born, Heisenberg, and Jordan in the matrix mechanics. The reader should take warning from this horrible example of the results of unconventionality!

ambiguity regard ing the order of the factors in the individual terms. We resolve this ambiguity by using in each case the mean value of the two products in direct and reverse order. This method of defining $O_{\mathfrak{S}}$ gives the corresponding matrix Hermitian form, and Eq. (45a) now follows as a corollary.

This completes the justification of Eq. (46) and indicates that one can make a certain amount of progress by the comparatively simple "quantization of the field" used in Eqs. (43) and (45a). The result is still incorrect, to be sure, since Eq. (47) cannot be derived from Eq. (46) by any legitimate procedure, but Frenkel's method does represent an advance over that employed on p. 14.

Consider next the phenomena of dispersion and scattering. Classically we account for them as due to the emission of secondary wavelets by atoms and molecules executing forced vibrations under the influence of the primary light waves. The first step in the theory is to compute the electric moment of the atom from the equations of motion of the electrons treated as oscillators held to fixed positions of equilibrium by elastic forces. Eqs. (43) are then used to compute the electric forces of the wavelets due to the individual atoms, and these are combined with due regard to amplitude and phase to give the scattered radiation and the coherently emitted beam which produces dispersion and reflection. In the quantum theory we may follow the same general program, the first step being the computation of the perturbed value of the mean electric moment of an individual atom under the influence of a monochromatic radiation field.

Let the field consist of a sinusoidal train moving along the x-axis with its initial wave-front at $x=ct$. We assume that the waves are polarized with the electric vector parallel to the y-axis, and define the vector potential by the equation

$$\begin{aligned} A_y &= -\alpha f(t-x/c) = -\alpha \sin 2\pi\nu_0(t-x/c), && t-x/c>0\,; \\ &= 0, && t-x/c<0. \end{aligned} \tag{54}$$

We may use the machinery of Section 4, 3 and Section 5, 3, neglecting as before the variation in the phase of the radiation field over the volume of the individual atom. Let ξ denote the x coordinate of the center of gravity of the atom or molecule under consideration. Replacing the fixed time T of Section 5, 3 by the variable time t of Eqs. (I, 178-179) we have in first approximation

$$c_m(t) = c_m{}^0 + \alpha \sum_k c_k{}^0 f_{km}^{(1)}(t) \tag{55}$$

where

$$f_{km}^{(1)}(t) = -\frac{4\pi^2}{hc}\nu_{km}P_y(km)\int_{t'=\xi/c}^{t'=t} f(t'-\xi/c)e^{2\pi i\nu_{km}t'}dt' \tag{56}$$

Expressing $f(t'-\xi/c)$ in terms of exponentials and carrying out the integration, the latter equation becomes

$$f_{km}(t)=\frac{\pi}{hc}\nu_{km}P_y(km)\left\{\frac{e^{2\pi i[(\nu_{km}+\nu_0)t-\nu_0\xi/c]}-e^{2\pi i\nu_{km}\xi/c}}{\nu_{km}+\nu_0}\right.$$
$$\left.-\frac{e^{2\pi i[(\nu_{km}-\nu_0)t+\nu_0\xi/c]}-e^{2\pi i\nu_{km}\xi/c}}{\nu_{km}-\nu_0}\right\} \tag{57}$$

provided that the radiation frequency ν_0 is not equal to the characteristic frequency ν_{km}, a condition which we shall suppose satisfied. Using the perturbed values of the c's given by Eqs. (55) and (57) in the general expression (38) for the mean electric moment of the atom, multiplying out, neglecting terms in the square of the amplitude of the light wave, and reducing, we obtain

$$\bar{P}(t)=\sum_{n,m}c_n{}^{0*}c_m{}^0\left\{\boldsymbol{P}(mn)e^{2\pi i\nu_{mn}t}\right.$$
$$+\frac{\alpha\pi}{hc}\sum_k\boldsymbol{P}(kn)P_y(mk)\nu_{mk}\left[\frac{e^{2\pi i[(\nu_{mn}+\nu_0)t-\nu_0\xi/c]}-e^{2\pi i(\nu_{kn}t+\nu_{mk}\xi/c)}}{\nu_{mk}+\nu_0}\right.$$
$$\left.-\frac{e^{2\pi i[(\nu_{mn}-\nu_0)t+\nu_0\xi/c]}-e^{2\pi i(\nu_{kn}t+\nu_{mk}\xi/c)}}{\nu_{mk}-\nu_0}\right]$$
$$+\frac{\alpha\pi}{hc}\sum_k\boldsymbol{P}^*(km)P_y{}^*(nk)\nu_{nk}\left[\frac{e^{2\pi i[(\nu_{mn}-\nu_0)t+\nu_0\xi/c]}-e^{2\pi i(\nu_{mk}t+\nu_{kn}\xi/c)}}{\nu_{nk}+\nu_0}\right.$$
$$\left.\left.-\frac{e^{2\pi i[(\nu_{mn}+\nu_0)t-\nu_0\xi/c]}-e^{2\pi i(\nu_{mk}t+\nu_{kn}\xi/c)}}{\nu_{nk}-\nu_0}\right]\right\}. \tag{58}$$

The coefficients of the initial amplitude products $c_n{}^{0*}\,c_m{}^0$ make up what we may call the perturbed value of the electric moment matrix $\boldsymbol{P}(m,\ n/t)$ to be used in computing $\boldsymbol{E}_{mn}$ and $\boldsymbol{H}_{mn}$. The terms having the frequencies ν_{mk} and ν_{kn} indicate a slight modification of the emission intensities due to the incidence of the perturbing light wave. In practice they would be quickly damped out by the radiation resistance which is always present, though we cannot include it in our theory, and we therefore neglect them.

There are three sorts of terms in the matrix, viz., those having the characteristic frequencies ν_{mn} of the unperturbed atom and giving rise to spontaneous incoherent emission; those having the frequency of the exciting light and producing reflection, dispersion, and normal Rayleigh scattering; and finally those having combination frequencies and producing Raman scattering. The Compton modified type of scattering is not represented in the above formula, as it does not take into account the continuous spectrum of characteristic functions of the atom. Compton scattering is entirely analogous to Raman scattering except that it is accompanied by atomic transitions from discrete energy levels to levels in the continuous spectrum.[29]

[29] In other words, the energy given up by a photon which experiences Compton scattering produces ionization of the atom struck, while the energy given up by a photon which experiences Raman scattering merely produces excitation of the atom struck.

The diagonal terms obtained by setting m equal to n are responsible for normal scattering and dispersion. If $\overline{D}(t)$ is used to denote the part of the mean electric moment coming from these diagonal terms, we have, after some juggling of terms,

$$\overline{D}(t)=\frac{2\alpha\pi}{hc}\sum_n |c_n^0|^2 R\left\{\sum_k \nu_{nk} P(kn) P_y(nk)\left[\frac{e^{2\pi i\nu_0\zeta}}{\nu_{nk}+\nu_0}-\frac{e^{-2\pi i\nu_0\zeta}}{\nu_{nk}-\nu_0}\right]\right\} \quad (59)$$

where the symbol R indicates that only the real part of the expression which follows it is to be considered, and where ζ stands for $(t-\xi/c)$.[30]

Although we have assumed a light wave polarized with the electric vector parallel to the y-axis, the above formula shows that the part of the induced electric moment having the frequency ν_0 is not rigorously parallel to the y-axis. In consequence the scattered light is not completely polarized. The high degree of polarization observed experimentally is accounted for by the fact that while $P_y(m, n)\, P_y(n, m)$ is equal to $|P_y(n, m)|^2$ and hence is essentially positive, such products as $P_x(m, n)\, P_y(n, m)$ may be either positive or negative, and will on the average tend to cancel each other when one carries out the summation indicated in Eq. (59).

The important y-component of $\overline{D}$ which is alone responsible for ordinary dispersion is readily reduced to the form

$$\overline{D}_y(t)=-\frac{4\alpha\pi\nu_0}{hc}\cos 2\pi i\nu_0\zeta \sum_n |c_n^0|^2\left\{\sum_k \nu_{nk} |P_y(nk)|^2\left(\frac{1}{\nu_{nk}^2-\nu_0^2}\right)\right\}. \quad (60)$$

To get the electric moment per unit volume we average $\overline{D}_y$ over all the atoms in a unit volume. Denoting the number of atoms per unit volume by N, we have

$$\begin{aligned} n^2-1 &= 4\pi\frac{\text{Electric moment per unit volume}}{\text{Primary electric force}} \\ &= \frac{8\pi N}{h}\sum_n |c_n^0|^2 \sum_m \nu_{mn} |P_y(nm)|^2\left(\frac{1}{\nu_{nm}^2-\nu_0^2}\right) \\ &= \frac{8\pi N}{h}\sum_{\nu_{nm}>0}\sum\left[|c_n^0|^2-|c_m^0|^2\right]|P_y(nm)|^2\frac{\nu_{nm}}{\nu_{nm}^2-\nu_0^2}. \end{aligned} \quad (61)$$

This is Kramer's dispersion formula. The appearance of $|c_m^0|^2$ with a negative sign shows that the atoms in the lower energy levels make contributions to dispersion opposite in sign to those in the upper energy levels.

Eq. (61) can be put in more conventional form by using our previous considerations (Section 5, 3) on degenerate systems. We replace the indices (n, m) by $(n, l; m, k)$ where l and k are the quantum numbers which determine the orientation of the atom in space; i.e., l and k

[30] In deriving the above formula one makes use of the fact that since the operator for $\boldsymbol{P}$ or for any one of its components is a real function of the coordinates, the matrix $\boldsymbol{P}(m, n)$ is Hermitian. The reader will observe that the matrix elements of $\boldsymbol{P}$ and P_y which appear explicitly in (59) are the unperturbed elements.

are the magnetic quantum numbers associated with the elementary states n and m. In the absence of any constant external fields we may write

$$\sum_{n,l} \left| c_{n,l}^0 \right|^2 \sum_{m,k} \nu_{mn} \left| P_y(n,l;m,k) \right|^2 \left(\frac{1}{\nu^2_{nm} - \nu_0^2} \right)$$
$$= \frac{1}{3} \sum_{n,l} \frac{1}{g_n} \left| c_n^0 \right|^2 \sum_{m,k} \nu_{mn} \left| P(n,l;m,k) \right|^2 \left(\frac{1}{\nu_{nm}^2 - \nu_0^2} \right) \quad (62)$$

as the ν_{nm}'s do not depend on the quantum numbers l and k, and also as[31]

$$\sum_{l,k} \left| P_y(n,l;m,k) \right|^2 = \tfrac{1}{3} \sum_{l,k} \left| P(n,l;m,k) \right|^2 \quad (63)$$

by symmetry. But by Eqs. (28) and (32)

$$\frac{1}{g_n} \sum_{l,k} \left| P(n,l;m,k) \right|^2 = \frac{3h^2}{8\pi^3} B_{n \to m} = \frac{3c^3 h}{64\pi^4 \nu_{nm}^3} A_{n \to m}. \quad (64)$$

Combining Eqs. (61), (62), and (64) we get

$$n^2 - 1 = \frac{Nc^3}{8\pi^3} \sum_n \left| c_n^0 \right|^2 \sum_m \frac{A_{n \to m}}{\nu_{nm}^2 (\nu_{nm}^2 - \nu_0^2)}$$
$$= \frac{Nc^3}{8\pi^3} \sum_{\nu_{nm} > 0} \sum \left[\left| c_n^0 \right|^2 - \left| c_m^0 \right|^2 \right] \frac{A_{n \to m}}{\nu_{nm}^2 (\nu_{nm}^2 - \nu_0^2)}. \quad (65)$$

The Rayleigh scattering for a gas can be computed from the induced electric moment per atom $\overline{\boldsymbol{D}}$ as in the classical theory by adding the intensities of the waves scattered by the different atoms. The justification for this procedure depends on the haphazard location of gas molecules, or on the fluctuations in gas density, according to one's point of view.

The Raman scattering comes from the non-diagonal terms in Eq. (58). Neglecting the terms involving only the characteristic frequencies of the atom, and the terms which produce normal scattering and dispersion, we may lump the others together and call the resultant the "Raman moment" $\boldsymbol{R}(t)$ of the atom. This gives

$$R_{mn}(t) = \frac{\alpha\pi}{hc} \left\{ \left[\sum_k \frac{\boldsymbol{P}(kn) P_y(mk) \nu_{mk}}{\nu_{mk} + \nu_0} + \frac{\boldsymbol{P}^*(km) P_y^*(nk) \nu_{kn}}{\nu_{nk} - \nu_0} \right] e^{2\pi i (\nu_0 + \nu_{mn}) t} \right.$$
$$\left. - \left[\sum_k \frac{\boldsymbol{P}(kn) P_y(mk) \nu_{mk}}{\nu_{mk} - \nu_0} + \frac{\boldsymbol{P}^*(km) P_y^*(nk) \nu_{kn}}{\nu_{nk} + \nu_0} \right] e^{-2\pi i (\nu_0 - \nu_{mn}) t} \right\}. \quad (66)$$

This is substantially the expression for a typical element of the Raman moment matrix of the atomic system as used elsewhere by the authors in discussing the Raman effect.[32]

If we compute the intensities of the various Raman lines by means of Eq. (66) we meet with the same difficulty as in dealing with spontaneous

[31] This formula is of considerable generality, and is intimately connected with the question of "spectroscopic stability." Eqs. (62) and (63) can be shown to be correct to a very high degree of approximation even in case one or more constant external fields are present. Cf. J. H. Van Vleck, Phys. Rev. **29**, 727 (1927); **30**, 31 (1927).

[32] E. L. Hill and E. C. Kemble, Proc. Nat. Acad. Sci. **15**, 387 (1929). In the equation used in this paper the anti-Stokes terms of the Raman moment (i.e. those with frequencies greater than that of the exciting light) were omitted, as were the frequency factors ν_{mk} and ν_{kn} in the numerators of the fractions in Eq. (66).

emission. The theory makes the intensities depend in a symmetrical way on the populations of the initial and final energy levels. To get an empirically correct formula one must first resolve the matrix $\{R_{mn}(\tau)\}$ into two parts $\{R_{mn}^{(-)}\}$ and $\{R_{mn}^{(-1)}\}$ associated with the anti-Stokes frequencies $(\nu_0+\nu_{mn})$ and the Stokes frequencies $(\nu_0-\nu_{mn})$ respectively. The emission from $\{R_{mn}^{(-)}\}$ is then computed using the populations of the upper energy levels, and from $\{R_{mn}^{(-1)}\}$ using the populations of the lower levels.

SECTION 6. MATRIX THEORY

I. Matrix Algebra and the Characteristic Value Problem

In Part I, Section 3, and in subsequent sections we have made use of certain quantities which we called "matrix elements" because they appear in sets, each of which can be arranged in a two-dimensional array, or "matrix." Each such set is derived from some "dynamical variable" α (a function of the coordinates and momenta of the problem) and from a complete normalized orthogonal set of functions of the q's, say $f_1, f_2, \cdots, f_n, \cdots$. In order to find the matrix[83] of α one must first set up a corresponding operator O_α (in simple cases by the substitution of $-ih/2\pi i\,\partial/\partial q_k$ for p_k in the classical formula for α) and then expand the function $O_\alpha f_m$ in terms of the complete set of f's. The matrix element $\alpha(m, n)$ is then defined by the expansion

$$O_\alpha f_m = \sum_n \alpha(m,n) f_n, \tag{67}$$

or by the corresponding explicit formula

$$\alpha(m,n) = \int_\infty f_n^* O_\alpha f_m d\tau. \tag{68}$$

We usually identify the f's with the characteristic functions which describe monochromatic solutions of Schrödinger's wave equation. We assume for the present that this equation has no continuous spectrum.[84] The above notation for the matrix elements will be used when f_n is set equal to the space factor $u_n(q)$, but if the complete characteristic function

$$\psi_n(q,t) = u_n(q) e^{2\pi i \beta_n t}$$

is to be used for f_n the symbol $\alpha(m,n)$ will be replaced by $\alpha(m, n; t)$.

[83] The matrix of α is said to be a "representation" of the operator O_α. It may be proved that the operator and the matrix are equivalent in the sense that either one determines the other. This is true despite the fact that the form of the matrix depends on the particular set of wave functions which is used. The equivalence depends on the invariant properties of the elements $\alpha(m, n; t)$ under a linear substitution of the functions ψ_n, a complete study of which would lead us far afield into the theory of linear substitutions which constitutes part of the mathematical theory of groups. Cf. A. Landé, "Optik, Mechanik, und Wellenmechanik," Sect. VII, Handbuch der Physik XX, Berlin, 1928.

[84] Every problem involving a continuous spectrum may be approximated by means of a modified problem in which the wave function is artificially required to vanish on some hypersphere of large radius in configuration space.

The matrices based on the characteristic functions may be used to determine the mean values of various dynamical variables in the case of an ensemble of systems whose ψ function is analysed into a sum of orthogonal characteristic functions. Several examples of this use have been given in the preceding section. It is also possible with the aid of matrices to reduce to a purely algebraic form much of the computation involved in solving problems in quantum mechanics. We have, in fact, shown (Part I, Sect. 4, 1) that if the matrix of the Hamiltonian operator is known for a complete set of normalized orthogonal functions of the coordinates, the location of the correct energy levels may be reduced to the algebraic problem of solving an infinite series of simultaneous linear equations in an infinite number of unknowns. (Cf. Eq. (I, 135)). A further study of this problem forms a natural method of approach to the matrix mechanics of Heisenberg, Born, and Jordan.

It will be convenient to begin with a brief review of the elements of matrix algebra. For this purpose we introduce the single symbol[35] $\boldsymbol{A}$ or $\{A(m,n)\}$ for the matrix

$$\boldsymbol{A} \equiv \{A(m,n)\} \equiv \begin{Vmatrix} A(11) & A(12) \cdots \\ A(21) & A(22) \cdots \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \end{Vmatrix}.$$

Two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be equal if, and only if, corresponding elements are equal; *i. e.*, if

$$A(m,n) = B(m,n). \qquad m,n = 1,2, \cdots, \infty.$$

To add or subtract two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we add or subtract corresponding elements.
Thus

$$(A \pm B)(m,n) = A(m,n) \pm B(m,n).$$

To multiply two matrices we use the rule[28]

$$(AB)(m,n) = \sum_k A(m,k)B(k,m). \tag{69}$$

On the basis of the above rule it can readily be shown that the associative law of ordinary algebra applies to the multiplication of matrices, although the commutative law does not. Thus in general

$$\boldsymbol{AB} \neq \boldsymbol{BA}.$$

The product of a matrix $\boldsymbol{A}$ and an ordinary number c is the matrix whose typical element is $cA(m, n)$.

[35] In the preceding sections we used bold face type to distinguish three-dimensional vectors from scalars. In the remainder of the article, however, we shall have no need for vectors and *shall reserve the bold face type notation for matrices.*

The unit matrix is defined by

$$I \equiv \{\delta_{mn}\} \equiv \begin{Vmatrix} 1 & 0 & 0 \cdots \\ 0 & 1 & 0 \cdots \\ 0 & 0 & 1 \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{Vmatrix}.$$

It follows at once that

$$AI = IA = A.$$

The inverse or reciprocal of a matrix A is defined, in case it exists, by the relation

$$AA^{-1} = A^{-1}A = I. \tag{70}$$

Positive and negative integral powers of A can now be defined by the rules

$$A^p = AAAAAA \cdots \text{ to } p \text{ factors}$$
$$A^{-p} = (A^{-1})^p$$
$$A^0 = I.$$

With the aid of these definitions we may finally establish the usual exponential rules

$$A^\mu A^\nu = A^{\mu+\nu}; \quad (A^\mu)^\nu = A^{\mu\nu}$$

for all positive and negative integral powers of A.

Consider now the problem of determining the energy levels of an atomic system. Let a set of approximate wave functions $\psi_1^0, \psi_2^0, \cdots, \psi_n^0, \cdots$ be given, and let the corresponding matrix for the Hamiltonian operator be $\{H^0(m, n/t)\}$. From the Schrödinger point of view the problem is to replace the approximate wave functions by exact wave functions $\psi_1, \psi_2, \cdots, \psi_n, \cdots$ which satisfy the differential equation,

$$H\psi_n = E_n\psi_n = \frac{h}{2\pi i}\frac{\partial \psi_n}{\partial t}, \tag{71}$$

and to locate the corresponding characteristic values of E. If, however, the original set of ψ's forms a complete orthogonal set, any new set is expressible in terms of the old by means of a suitable linear expansion. Hence the replacement of the initial functions by the actual characteristic functions of the problem may be described as a *linear transformation* of the initial set of functions. This transformation has the form

$$\psi_m{}^0 \rightarrow \psi_m = \sum_k S(m, k/t)\psi_k{}^0. \tag{72}$$

the coefficients forming a matrix $\{S(m, n/t)\}$ which completely determines the transformation. If the new functions, like the old, form a normalized

orthogonal set, such as can always be picked out from the complete manifold of the solutions of Eq. (71), the elements of the transformation matrix satisfy the relation[36]

$$\sum_k S(n,k/t)S^*(m,k/t)=\delta_{nm}. \tag{73}$$

Eq. (73) is a generalization of the relations between the direction cosines in a rotation of axes in three-dimensional space. A linear transformation which satisfies (73) is said to be an *unitary* transformation.[37]

If we introduce the symbol $\tilde{S}$ for the matrix obtained from S by interchanging rows and columns, we see that the condition (73) is equivalent to the matrix equation

$$\tilde{S}^*S=S\tilde{S}^*=I. \tag{74}$$

But by the definition of the inverse matrix S^{-1} given in Eq. (70) this is equivalent to writing

$$S^*=S^{-1}. \tag{75}$$

Turning now to the matrix of the Hamiltonian operator, we readily prove that the application of the transformation whose matrix is S to the ψ functions leads to a similar linear transformation of the elements of the matrix of H. Thus

$$\begin{aligned} H^0(m,n/t)\rightarrow H(m,n/t) &= \int \psi_n^* H\psi_m d\tau \\ &= \int \left[\sum_l S^*(n,l/t)(\psi_l^0)^*\right] H\left[\sum_k S(m,k/t)\psi_k^0\right] d\tau \\ &= \sum_{l,k} S^*(n,l/t)H^0(k,l/t)S(m,k/t). \end{aligned} \tag{76}$$

In the language of matrix algebra, Eq. (76) takes the form

$$H^0\rightarrow H=SH^0\tilde{S}^*=SH^0S^{-1}. \tag{77}$$

Equation (77) describes the transformation of the Hamiltonian matrix due to the substitution of any new complete orthogonal set of functions for the initial set. We wish to demand, however, that the new set shall be the actual characteristic functions of the problem in hand. In that case we have directly by Eq. (71) that

$$H(m,n/t)=\int \psi_n^* E_m\psi_m d\tau = E_m\delta_{nm}. \tag{78}$$

In other words, we require the transformation to carry the matrix of H over into what is called *diagonal* form, all terms dropping out except those

[36] Cf. section in fine print at top of p. 211, Part I; also Eq. (I, 182).

[37] This is a slight generalization of the *orthogonal* transformation adapted to the algebra of complex quantities. Cf. H. Weyl, "Gruppentheorie und Quantenmechanik," p. 16, Leipzig, 1928.

for which $m=n$. The diagonal terms are then the desired energy values. The problem of locating the exact energy levels and exact wave functions is thus reduced to the problem of finding a matrix $\boldsymbol{S}$ which will carry $\boldsymbol{H}$ over into diagonal form when applied according to the rule of Eq. (77). If we now equate the right hand member of Eq. (77) to the matrix $\{E_m\delta_{mn}\}$, and multiply both sides of the equation behind by the matrix $\boldsymbol{S}$, we obtain[38]

$$\boldsymbol{S}\boldsymbol{H}^0=\{E_m\delta_{mn}\}\boldsymbol{S}. \tag{79}$$

Equating the two members of this equation element by element, we obtain a double infinity of simultaneous equations with the elements of the matrix $\boldsymbol{S}$ and the E values as unknowns; i.e.,

$$\sum_k S(m,k/t)[H^0(k,n/t)-E_m\delta_{kn}]=0, \quad m,n=1,2,\cdots\infty. \tag{80}$$

If the initial set of approximate wave functions are of the form

$$\psi_m{}^0=u_m{}^0e^{2\pi iE_m{}^0t/h} \tag{81}$$

where the u^0's are independent of the time, we can simplify the set of Eqs. (80). By Eq. (71) ψ_m is of the form $u_me^{2\pi iE_mt/h}$. Hence Eq. (72) is equivalent to

$$u_me^{2\pi iE_mt/h}=\sum_k S(m,k/t)u_k{}^0e^{2\pi iE_k{}^0t/h}.$$

It follows that

$$u_m=\sum_k S(m,k)u_k{}^0$$

where $S(m,k)$ is independent of t and

$$S(m,k/t)=S(m,k)e^{2\pi i(E_m-E_k{}^0)t/h}. \tag{82}$$

Similarly

$$H^0(m,k/t)=H^0(m,k)e^{2\pi i(E_m{}^0-E_k{}^0)t/h}. \tag{83}$$

By means of Eqs. (82) and (83), Eqs. (80) are readily reduced to the time-free form

$$\sum_k S(m,k)[H^0(kn)-E_m\delta_{kn}]=0, \quad m,n=1,2,\cdots,\infty. \tag{84}$$

The set of equations obtained from the above by picking out all cases in which m has a given value is identical with the set (I, 135) except for the detail of notation. As explained in Part I, the solution of these equations can be carried through in general only by the successive approximations of perturbation theory.

So far we have used matrices only to re-state, in a different language, conclusions already reached in Part I by formally different methods. We can go much further than this, however, with the aid of the theorem of page 18 which states that the matrix of the product of two operators in terms of

[38] Cf. Born, Heisenberg, and Jordan, Zeits. f. Physik **35**, 557 (1926), Eq. (20).

any given set of orthogonal functions is equal to the product in inverse order of the matrices of the two operators taken independently. Since the H operator is built up from the operators for the p's and q's, it follows at once that the H matrix can be built up from the matrices of the p's and q's. Similarly we can build up the matrix of the electric moment from the matrices of the p's and q's. Thus the solution of both the energy level problem and the intensity problem can be made to turn on the determination of the matrices for the p's and q's. In what follows we shall see that in some cases, at least, the matrices in question can be found and the whole problem solved without the specific introduction of the initial orthogonal function system $\psi_1^0, \psi_2^0, \cdots$.

2. The Properties of the p and q Matrices

The matrices for the coordinates q_k and the momenta p_k for a Cartesian coordinate system are derived from the operators q_k and $-(h/2\pi i)\partial/\partial q_k$ respectively. In conformity with the last paragraph of Section 3, 4 of Part I we provisionally define the operator for the momentum conjugate to the *generalized* coordinate q_k to be $-(h/2\pi i)\partial/\partial q_k$.[38a] Hence for any conjugate variables q_k and p_k[28]

$$[p_k q_k - q_k p_k](m,n) = \int \psi_n^* \left[q_k \left(-\frac{h}{2\pi i}\frac{\partial}{\partial q_k} \right) + \frac{h}{2\pi i}\frac{\partial}{\partial q_k} q_k \right] \psi_m d\tau$$

$$= \int \psi_n^* \frac{h}{2\pi i} \psi_m d\tau = \frac{h}{2\pi i}\delta_{mn}. \tag{85}$$

This is equivalent to the matrix equation

$$p_k q_k - q_k p_k = (h/2\pi i)I. \tag{86}$$

By a similar argument we can prove that

$$p_k q_l - q_l p_k = 0, \quad l \neq k \tag{87}$$

and

$$p_k p_l - p_l p_k = q_k q_l - q_l q_k = 0. \tag{88}$$

In other words, the matrices for coordinates do not "commute" with the matrices for the conjugate momenta, though they do commute with the matrices for the other momenta. Also any momentum matrix commutes with any other, and any coordinate matrix with any other.

The fundamental relation of Eq. (86) between the matrices of the coordinates and of the conjugate momenta is the quantum mechanical equivalent of the Wilson-Sommerfeld quantum conditions of the Bohr theory[39]

[38a] Cf. pp. 36 and 42. The differential operator does not apply however when the coordinates assume discrete values as for instance in the use of a component of angular momentum as a matrix. Cf. also Sect. 7, 2 below.

[39] M. Born and P. Jordan, Zeits. f. Physik **34**, 858 (1925). Cf. also L. Brillouin, Journ. de Physique **7**, 135 (1926); reprinted in English in the volume "Selected Papers on Wave

$$\int p_k dq_k = nh \quad n = 0, 1, 2, \cdots .$$

The Eqs. (86), (87), (88) are called the "commutation rules." Dirac introduces the Poisson bracket symbol $[y, x]$ for the matrix function $(2\pi i/h)$ $(xy - yx)$, and writes the commutation rules in the form[40]

$$[q_k, p_l] = I\delta_{kl}, \quad [q_k, q_l] = [p_k, p_l] = 0. \tag{89}$$

Another fundamental characteristic of the matrices of the p's and q's in Cartesian coordinates is their *Hermitian character* (Cf. Part I, p. 201) which is a consequence of the fact that the corresponding operators are *self-adjoint.* An operator O_α is by definition self-adjoint with respect to a coordinate system $q_1, \ldots, q_{3n}$, if

$$vO_\alpha u - uO_\alpha^* v = \sum_{k=1}^{3n} \frac{\partial f_k}{\partial q_k}, \tag{90}$$

where u and v are arbitrary continuous functions of the coordinates $q_1, \cdots, q_{3n}$, and $f_1, \cdots, f_{3n}$ are functions of the same coordinates, any or all of which may vanish in special cases.[40a] (The right hand member of the above equation is simply the divergence of the $3n$-dimensional vector F whose components are $f_1, \cdots, f_{3n}$). If we now take the volume integral of both sides of Eq. (90) over a hyper-sphere G surrounded by a surface S, we can convert the right hand member into a surface integral over S. If the functions u and v vanish rapidly at infinity, the vector F approaches zero with corresponding rapidity, and the surface integral vanishes in the limit as the radius of the sphere increases indefinitely. Under these circumstances

$$\int_\infty (vO_\alpha u - uO_\alpha^* v) dq_1 \cdots dq_{3n} = 0. \tag{91}$$

If u and v are identified respectively with the wave function ψ and its conjugate ψ^* for a suitable problem in wave mechanics we have

$$\bar{\alpha} = \int_\infty \psi^* O_\alpha \psi d\tau = \int_\infty \psi O_\alpha^* \psi^* d\tau = \bar{\alpha}^*. \tag{92}$$

Thus the mean value of a quantity α with a self-adjoint operator O_α is necessarily real provided that the ψ function for the problem in hand vanishes

---

Mechanics" by de Broglie and Brillouin, London, 1928. It may be remarked incidentally that Eqs. (86) can only be satisfied if q_k and p_k are infinite matrices. Cf. e.g., H. Turnbull, Mathematical Gazette **14,** 12 (1928).

[40] Dirac's notation is dictated by the fact that in the limiting case of large quantum numbers the matrix quantity $(2\pi i/h)$ $(xy - yx)$ corresponds to the classical Poisson bracket $[y,x]$ just as the matrix p corresponds to the classical value of the momentum p. Cf. P. A. M. Dirac, Proc. Roy. Soc. **A109,** 642 (1925).

[40a] An alternative definition of the self-adjoint operator equivalent to Eq. (90) is given by Jordan [Zeits. f. Physik **40,** 809 (1927)]. Eq. (91) can also be used to define the self-adjoint property.

with sufficient rapidity at infinity.[41] The latter restriction is of greater importance from the mathematical than from the physical point of view.

We can also identify u and v^* respectively with two characteristic functions ψ_n and ψ_m belonging to the discrete spectrum of the problem in hand. Then

$$\int_{\infty} (\psi_m^* O_\alpha \psi_n - \psi_n O_\alpha^* \psi_m^*) d\tau = 0$$

or

$$\alpha(n, m/t) = \alpha^*(m, n/t), \tag{93}$$

which is the condition that the matrix of α shall be Hermitian.

In classical mechanics and electrodynamics complex values of the dynamical variables have no meaning, and in quantum mechanics we may reasonably exclude complex values of measurable quantities as meaningless. This is equivalent to requiring that the operators correlated with all measurable quantities shall be self-adjoint, and that the corresponding matrices shall all be Hermitian. We have already proved that these conditions are satisfied by the energy operator and energy matrix. They are also satisfied for all functions of the coordinates only, and for the components of linear and angular momentum, as the reader can prove for himself without difficulty. Hereafter we shall assume that only self-adjoint operators and Hermitian matrices are to be used in building up the theory.

We have now proved that in order to solve the characteristic value problem for a given Hamiltonian operator H we can adopt as a starting point any complete orthogonal function system spread out over the whole of the configuration space for the problem, develop from it matrices for the p's and q's, compound them to form a matrix for H, and then apply a transformation of the *canonical type*[42]

$$T_c : \left\{ \begin{array}{l} \boldsymbol{p}_k^0 \rightarrow \boldsymbol{p}_k = S \boldsymbol{p}_k^0 S^{-1} \\ \boldsymbol{q}_k^0 \rightarrow \boldsymbol{q}_k = S \boldsymbol{q}_k^0 S^{-1} \\ H^0 \rightarrow H = S H^0 S^{-1} \end{array} \right\} \quad S^{-1} = \tilde{S}^* \tag{94}$$

which reduces the Hamiltonian matrix to diagonal form. The initial matrices $(\boldsymbol{p}_k^0, \boldsymbol{q}_k^0)$ must be Hermitian and must satisfy the commutation rules. Without appreciable loss of generality we may assume that they have the standard form

$$p^0(m, n/t) = p^0(m, n) e^{2\pi i (E_m^0 - E_n^0) t/h}. \tag{95}$$

[41] It seems probable that the existence of the integral which defines $\bar{\alpha}$ is a sufficient practical criterion for the validity of Eq. (92).

[42] The term "*canonical*" is applied to this type of transformation because it preserves the commutation rules and hence the quantum mechanical form of Hamilton's equations of motion (to be derived below). The connection between this form of canonical transformation and the form used in the classical Hamiltonian theory has been traced by Jordan. Cf. P. Jordan, Zeits. f. Physik **38**, 513 (1926).

For the rest, it is clear that a wide variety of forms must be possible in any given case, and one surmises that any choice consistent with the above restrictions is permissible. In other words, it would not be unreasonable to assume—and in fact it has been commonly assumed—that the problem of finding a set of Hermitian matrices for the p's and q's which obey the commutation rules and reduce $\boldsymbol{H}(\boldsymbol{p}, \boldsymbol{q})$ to diagonal form has no other solutions than those obtainable from the Schrödinger equation. In the case of a non-degenerate system this would mean that the solutions of the matrix problem are unique.

In practice the solution of problems in quantum mechanics by starting from an arbitrary initial set of Hermitian matrices which satisfy the commutation rules, and applying to them a canonical transformation which reduces $\boldsymbol{H}(\boldsymbol{p}, \boldsymbol{q})$ to diagonal form is confined to cases in which the Hamiltonian function can be approximated by that of a slightly different problem which has been exactly solved and whose matrices form the starting point of a genuine perturbation method calculation.

3. The Equations of Motion in the Matrix Theory

The original formulation of the matrix mechanics of Born, Heisenberg, and Jordan was based on an attempt to use, so far as possible, the formal machinery of the classical dynamics, but with a reinterpretation of the nature and the meaning of the symbols introduced. They *assumed* that the coordinates and momenta of every dynamical system were to be represented by matrices, that the Hamiltonian function was a function of the coordinate and momentum matrices, and that the canonical equations of Hamilton were applicable to these matrices. On the basis of these assumptions they were able to prove that the energy matrix $\boldsymbol{H}(\boldsymbol{p}, \boldsymbol{q})$ was constant in time (law of the conservation of energy),[43,44] which they showed to be equivalent to the statement that the energy matrix was of diagonal form if the system was non-degenerate. Otherwise they found it necessary to *assume* explicitly the diagonal form of $\boldsymbol{H}$.

According to our present point of view, the matrices of p, q, and H can have a great variety of equally legitimate forms since the ψ functions can be described in terms of many different orthogonal function systems. The solution of the characteristic value problem is obtained, however, by reducing the matrices to the form which corresponds to an expansion in terms of a complete orthogonal system of characteristic functions. We may call this the normal matrix form. By definition $\boldsymbol{H}(\boldsymbol{p}, \boldsymbol{q})$ is constant in time and diagonal if the matrices of the p's and q's are in normal form. We may now reverse the argument of Born and Heisenberg to prove that normal matrices satisfy equations formally identical with the Hamiltonian equations of classical theory. As a first step we must define the operation of differentiation of a matrix function.

[43] Cf. M. Born and P. Jordan, Zeits. f. Physik **34**, 858 (1925).

[44] M. Born, W. Heisenberg, and P. Jordan, Zeits. f. Physik **35**, 557 (1926).

Born, Heisenberg, and Jordan give two definitions of the derivative of a matrix function with respect to a matrix argument.[44] We shall use here what they call "derivatives of the first kind," which are defined by the equation

$$\frac{df}{dx}=\lim_{a\to 0}\frac{f(x+a)-f(x)}{a} \tag{96}$$

where a is the diagonal matrix obtained by multiplying the unit matrix I by an ordinary number a. It is readily proved that on the basis of this definition all the ordinary rules of differentiation apply to the differentiation of matrix functions except that one must have due regard for the order in dealing with product functions. Thus

$$\frac{d}{dx}(FG)=F\frac{dG}{dx}+\frac{dF}{dx}G,$$

where the order of the products cannot be altered in either member unless it is reversed throughout the equation.

It follows from this definition that if $f(p, q)$ is any function of a set of matrices (q_k, p_k) which satisfy the commutation rules[45]

$$\begin{aligned}\frac{\partial f}{\partial p_k}&=(2\pi i/h)(fq_k-q_kf)=[q_k,f]\\ \frac{\partial f}{\partial q_k}&=(2\pi i/h)(p_kf-fp_k)=[f,p_k].\end{aligned} \tag{97}$$

Differentiation of a matrix f with respect to an ordinary number a is defined by the rule

$$\left(\frac{df}{da}\right)(m,n)=\frac{d}{da}f(m,n). \tag{98}$$

Applying the first of equations (97) to the Hamiltonian matrix $H(p_k, q_k)$ we have

$$\frac{\partial H}{\partial p_k}=\frac{2\pi i}{h}(Hq_k-q_kH). \tag{99}$$

If the p's and q's are normal matrices, H is a diagonal matrix whose non-vanishing elements are the energy levels $E_1, E_2, \cdots$. Hence

$$\left(\frac{\partial H}{\partial p_k}\right)(m,n)=\sum_j\frac{2\pi i}{h}\{E_m\delta_{mj}q_k(j,n/t)-q_k(m,j/t)E_n\delta_{nj}\}$$

[45] The proof of this theorem rests on the observation that it is true if we identify f with any one of the basic matrices $q_1, q_2, \cdots, p_1, p_2, \cdots$. It is easy to show that if the theorem is true for any pair of functions F and G it is true for their sum and for their product. If we restrict ourselves to matrix functions which can be expressed as the sums of products of integral powers of the p's and q's, the rest follows by induction.

$$= \frac{2\pi i}{h}(E_m - E_n)q_k(m,n/t) \tag{100}$$

$$= \frac{d}{dt}q_k(m,n/t).$$

It follows from Eq. (98) that

$$\frac{dq_k}{dt} = \frac{\partial H}{\partial p_k}. \tag{101}$$

The second Hamiltonian equation

$$\frac{dp_k}{dt} = -\frac{\partial H}{\partial q_k} \tag{102}$$

is proved in exactly the same way.

Once Eqs. (101) and (102) have been established for normal matrices there is no difficulty in showing that they remain valid for any set of matrices obtained from normal ones by a canonical transformation with a constant matrix S.[46]

The equations of motion (101) and (102) were of considerable importance in the initial development of the matrix theory, since their direct integration in a few special cases gave clean solutions of simple problems which could then be used as the basis for an attack on more difficult problems by perturbation methods.[47]

4. Example: The Linear Oscillator

Consider the case of the linear oscillator whose Hamiltonian function has the form

$$H(p,q) = p^2/2\mu + kq^2/2. \tag{103}$$

The Hamiltonian equations of motion reduce to

$$\dot{q} = p/\mu \; ; \quad \dot{p} = -kq. \tag{104}$$

[46] The proof rests on the theorem that if

$$p_k = Sp_k{}^0S^{-1}; \quad q_k = Sq_k{}^0S^{-1}$$

then

$$f(p_k, q_k) = Sf(p_k{}^0, q_k{}^0)S^{-1}$$

and on the additional fact that if S is a constant matrix

$$dp_k/dt = S(dp_k{}^0/dt)S^{-1}.$$

The transformed matrices obtained from normal matrices by means of a constant canonical transformation do not have the form of Eq. (95).

[47] For examples of papers dealing with problems on the matrix (or q-number) theory which use the equations of motion, the following references can be consulted. In the papers of Mensing and Oppenheimer a certain amount of perturbation theory is used. P. A. M. Dirac, Proc. Roy. Soc. **A110,** 561 (1926), W. Pauli, Jr., Zeits. f. Physik **36,** 336 (1926), L. Mensing, Zeits. f. Physik **36,** 814 (1926), J. R. Oppenheimer, Proc. Cam. Phil. Soc. **23,** 327 (1926), I. Tamm, Zeits, f. Physik **37,** 685 (1926), D. M. Dennison, Phys. Rev. **28,** 318 (1926).

Eliminating p and introducing the classical frequency $\nu_0[=(1/2\pi)\ (k/\mu)^{1/2}]$ we have

$$\ddot{q}=\{(2\pi i\nu_{mn})^2 q(m,n/t)\}=-(2\pi\nu_0)^2 q$$

or

$$[\nu_{mn}^2-\nu_0^2]q(m,n/t)=0. \tag{105}$$

This equation shows that $q(m, n/t)$ must vanish unless

$$\nu_{mn}=\frac{E_m-E_n}{h}=\pm\nu_0. \tag{106}$$

It follows that the energy levels, or diagonal elements of the H matrix, are arranged in one or more equidistant series, and that there are just two non-vanishing elements in each row and column. From the Hermitian character of the matrices we know, of course, that the non-vanishing elements are symmetrically arranged with respect to the diagonal. Let us now assume that there is only one set of equidistant energy values as in the Bohr theory. The order of the elements of a diagonal matrix is arbitrary, so that we can assume without loss of generality that it is the same as the order of the energies. The non-vanishing elements of the matrices for q and p are then arranged in two lines parallel and adjacent to the principal diagonal. It follows at once that all non-diagonal elements of $[q, p]$ vanish as required by the commutation rule. The diagonal elements of $[q, p]$ must have the common value unity. In other words

$$\frac{h}{8\pi^2\mu}=-\sum_k \nu_{nk}\,|\,q(n,k/t)\,|^2=\nu_0[\,|\,q(n,n+1/t)\,|^2-|\,q(n;n-1/t)\,|^2]. \tag{107}$$

Hence the squares of the absolute values of the matrix elements $q(n, n+1/t)$ form an arithmetical progression with the common difference $h/8\pi^2\mu\nu_0$. These terms are essentially positive so that n must have a minimum value which we shall set equal to zero. For $n=0$, Eq. (107) then reduces to

$$|\,q(0,1/t)\,|^2=h/8\pi^2\mu\nu_0$$

and we have finally

$$|\,q(n,n+1/t)\,|^2=(n+1)h/8\pi^2\mu\nu_0,\quad n=0,1,2,\cdots. \tag{108}$$

For the matrix element $q(n+1, n/t)$ this yields

$$q(n+1,n/t)=q^*(n,n+1/t)=\left[\frac{(n+1)h}{8\pi^2\mu\nu_0}\right]^{1/2} e^{i(2\pi\nu_0 t+\phi_n)}, \tag{109}$$

where ϕ_n is an arbitrary phase constant. The factors $e^{i\phi_n}$ may be lumped with the complex amplitudes c_n in determining mean values in the case of a particular ensemble of linear oscillators. [Cf. for example, Eq. (38).] In the simple problem under consideration the electric moment is proportional to the displacement q so that by combining Eq. (108) with Eqs.

(28) and (35) we obtain at once the transition probabilities with which to compute the intensity of the emitted radiation.

Having evaluated the q matrix, we can next determine p by Eq. (104) and H by Eq. (103). For the latter we find

$$H(n,n/t)=E_n=(n+\tfrac{1}{2})h\nu_0. \tag{110}$$

These conclusions are in agreement with the pure wave theory of Sect. 2, 5 of Part I, and we thus see that the two theories lead to equivalent results.[48]

5. Hamilton's Equation for Wave Packets

Due to the general relation between matrices and mean values, viz.,

$$\bar{\alpha}=\int\psi^*O_\alpha\psi d\tau=\sum_{n,m}c_n^*c_m\alpha(m,n/t)$$

it follows that all matrix equations such as the Hamiltonian Eqs. (101) and (102) are equally valid if we replace each member by the corresponding mean value for some definite ensemble of systems. Thus we have

$$\overline{\frac{dq_k}{dt}}=\frac{d}{dt}\overline{q_k}=\overline{\frac{\partial H}{\partial p_k}};\quad \overline{\frac{dp_k}{dt}}=\frac{d}{dt}\overline{p_k}=-\overline{\frac{\partial H}{\partial q_k}}. \tag{111}$$

If, in particular, we consider an ensemble represented by a wave packet in which each of the coordinates and momenta is well defined, the mean value of any function of the coordinates and momenta will closely approximate the same function of the mean values of the individual coordinates and momenta. The classical mechanics is the mechanics of just such wave packets and the classical coordinates and momenta are evidently to be identified with the mean values of the actual coordinates and momenta. Thus in the limiting case where the coordinates and momenta are very well defined, Eqs. (111) go over into the classical Hamiltonian equations. The proof is independent of the coordinate system employed, and hence the theorem shows that the definition of the momentum operator in generalized coordinates given on p. 29 is justified by agreement with the classical definition of canonical pairs of variables in the limiting case of a sharply defined wave packet. On the other hand the proof does depend on the assumption that the Hamiltonian function has only a discrete spectrum of characteristic wave functions. As previously explained, one can approximate the solution of any problem in quantum mechanics involving a continuous spectrum by means of a modified

[48] Cf. in this connection L. Brillouin, Journ. de Physique **7**, 135 (1926). Brillouin's conclusion that the matrix theory leads to different results if n is allowed to assume negative values seems to the authors to be due to an algebraic error in his final energy formula. On checking the steps we find that the energy and coordinate matrices for his alternative solution are given by (110) and (109) with $(n+\frac{1}{2})$ and $(n+1)$ replaced by $(-n+\frac{1}{2})$ and $(-n+1)$ respectively, with $n=0,\ -1,\ -2,\ \cdots$. This amounts simply to a different system of numbering for the rows and columns and leads to no new results.

problem having only a discrete spectrum. Thus the continuous spectrum problem is a limiting case to which the Hamiltonian equations must also apply. A more direct treatment of the continuous spectrum case is given, however, in Sect. 7, 2.

6. Matrices and Bohr's Correspondence Principle

The matrix theory of Heisenberg, Born, and Jordan was developed in the beginning from Bohr's famous correspondence principle for conditionally periodic motions. This principle postulates the asymptotic identity of the frequencies and intensities of spectrum lines as observed and as computed from classical theory in the limiting case of high quantum numbers. The quantum mechanics is, of course, in harmony with Bohr's principle, which we have already used in a somewhat modified form in setting up the Schrödinger equation in Part I. As the process of tracing out the assumed correspondence is of practical value as well as historical interest, we briefly describe it here.

A conditionally periodic motion is a motion which from the classical standpoint should emit a line spectrum. In other words, it is a motion of such a character that every function of the coordinates and momenta, including the all-important electric moment, can be analysed into a multiple Fourier series. Thus if $P^{(x)}$ denotes the x component of the electric moment

$$P^{(x)}{}_{\text{classical}} = \sum_{\tau_1,\cdots,\tau_f=-\infty}^{+\infty} P^{(x)}_{\tau_1,\cdots,\tau_f} e^{2\pi i[\tau_1\nu_1+\cdots+\tau_f\nu_f]t} \tag{112}$$

where the τ's are positive or negative integers and the ν's are a set of fundamental frequencies characteristic of the motion. $P^{(x)}{}_{\tau_1,\cdots,\tau_f}$ is the complex conjugate of $P^{(x)}{}_{-\tau_1,\cdots-\tau_f}$ so that $P^{(x)}$ is real. A similar expansion holds for each individual coordinate and momentum component. The simplest type of conditionally periodic motion, and the one of the greatest importance in the Bohr theory, is that which occurs when by a suitable choice of coordinates the action function W can be resolved into the sum of a series of terms W_1, W_2, etc., each of which depends on a single coordinate. The variables are then said to be *separable*, and the Wilson-Sommerfeld quantum conditions can be applied to each of the coordinates, or what amounts to the same thing, to each of the terms in W. Under these circumstances the frequencies ν_1, ν_2, ν_3, $\cdots$ in Eq. (112) are the frequencies associated with the various coordinates used to define the system. Furthermore, each "stationary state," as defined by the quantum conditions, is associated with a set of quantum numbers n_1, n_2, $\cdots$ correlated in a one to one manner with the coordinates and with the frequencies.

Bohr has shown that according to the fundamental assumptions of his theory the frequency emitted in any jump, say

$$E(n_1', n_2', \cdots, n_f') \rightarrow E(n_1'', n_2'', \cdots, n_f''), \tag{113}$$

is equal to the mean value of the frequency $|(\tau_1\nu_1+\tau_2\nu_2+\cdots+\tau_f\nu_f)|$ of a corresponding pair of terms in Eq. (112) evaluated by taking a suitable

average over a continuous series of classically permissible, but unquantized, states between the initial and the final stationary states.[49] One of the corresponding terms is that for which $\tau_1 = n_1' - n_1''$, $\tau_2 = n_2' - n_2''$, etc., while the other is the conjugate term. In the limiting case of very high quantum numbers the classical frequencies ν_1, ν_2, ν_3, etc. have the same values for the initial and final states, and hence the quantum frequencies become actually identical with the corresponding classical combination frequencies.

In addition to this limiting agreement in the frequencies, which can be deduced from the quantum conditions and from his other quantum postulates, Bohr assumed a limiting agreement between the intensities of the terms in (112) and the intensities of the corresponding spectrum lines, and then used this assumption as a tool for the approximate computation of quantum intensities.

As a first step toward exhibiting the correspondence principle in terms of the quantum mechanics, we observe that since the classical action function W obeys a differential equation which is a limiting case of the differential equation for[50] $-(h/2\pi i)\log\psi$, it follows that the process of resolving ψ into the product of several factors is equivalent to resolving W into a sum of corresponding terms. Thus if the variables in the Schrödinger equation for ψ are separable in the sense of Section 2,6, Part I, it follows that the variables in the corresponding classical action function will be separable in the sense just defined.

Let us assume that this condition is fulfilled. Then in the wave mechanics there will be a differential equation for every coordinate, which can be thrown into Sturm-Liouville form, and each energy level will be characterized by a set of quantum numbers $n_1, n_2, \cdots, n_f$, one for each coordinate. We expect that just as in the Bohr theory the frequency and intensity of the radiation produced by the transition (113) will be asymptotically equal to the frequency and intensity of the corresponding term in Eq. (112).

Since the classical mechanics may be regarded as a limiting case of the quantum mechanics of wave packets when the coordinates and momenta of the latter are defined with precision, we can prove the point by writing out the wave mechanical equivalent of Eq. (112) for a well defined packet and actually passing to the limit. The necessary equation is evidently

$$\overline{P^{(x)}} = \sum_{n_1', \cdots, n_f'} \sum_{n_1'', \cdots, n_f''} c_{n_1', \cdots, n_f'} \overset{*}{c}_{n_1'', \cdots, n_f''} P^{(x)}(n_1', \cdots, n_f'; n_1'', \cdots, n_f'') e^{2\pi i \nu(n_1', \cdots, n_f'; n_1'', \cdots, n_f'')t} \tag{114}$$

This is equivalent to Eq. (38), but the notation has been amplified to bring out the explicit relation of each state to the corresponding quantum numbers. With the aid of suitable approximations the double sum in the above equation must collapse into a single sum like the right hand member of Eq. (112).

[49] Cf. J. H. Van Vleck, "Quantum Principles and Line Spectra" (Eq. (15a)) for an explicit statement of this theorem, which is also implicitly contained in Bohr's Danish Academy papers.

[50] Cf. Sect. 5, 2.

To accomplish the reduction the following assumptions are necessary:[51]

a) The characteristic functions in the packet which have amplitudes differing appreciably from zero all have very short wave-lengths. In other words, we are in the region of high quantum numbers. Without such functions it is impossible to fix coordinates and momenta simultaneously with the desired precision. (Cf. Part I, Section 1, 8 and 9).

b) The number of characteristic functions in the packet having amplitudes differing appreciably from zero must be very large if the mutual interference between these functions is to make the packet very small.

c) The range of energy values with appreciable amplitude coefficients must be so narrow that in this range the energy can be treated as a linear function of all of the quantum numbers.

d) The matrix elements of $P^{(x)}$ corresponding to transitions between the different characteristic functions of the packet must depend primarily on the quantum number differences between the different states involved, being practically independent of the absolute values of the initial quantum numbers through the range of the packet. This hypothesis is easily verified in special cases, and the work of Eckart and of Van Vleck[52] on the asymptotic relation between the matrix elements $P^{(x)}(n_1', \cdots, n_f'; n_1'', \cdots, n_f'')$ and the Fourier coefficients in Eq. (112) affords a sufficiently general proof.

e) Finally we must suppose that the amplitudes $c_{n_1,\cdots,n_f}$ which define the packet vary quite slowly in absolute value with the quantum numbers and have phases which like the energy values are linear functions of the quantum numbers. Then if the corresponding indices n_1', n_1''; n_2', n_2''; etc. differ by small integral values

$$c_{n_1',\cdots,n_f'} \cong c_{n_1'',\cdots,n_f''} e^{i[(n_1'-n_1'')\epsilon_1+(n_2'-n_2'')\epsilon_2+\cdots]}.$$

We shall undertake no careful verification of the above assumptions but shall content ourselves with exhibiting the relation between matrix elements and classical Fourier components to which they lead. For a more complete treatment of the problem the reader is referred to the paper by Van Vleck cited in footnote 52.

Let

$$\tau_k = n_k' - n_k''; \quad k = 1, 2, 3, \cdots, f.$$

[51] The writers are indebted to Professor H. A. Kramers for assistance in setting up the following scheme of reduction.

[52] C. Eckart, Proc. Nat. Acad. Sci. **12**, 684 (1926). J. H. Van Vleck, Proc. Nat. Acad. Sci. **14**, 178 (1928). Strictly speaking, there is an arbitrariness in the values of the matrix elements due to the arbitrariness in the phases of the characteristic functions. Thus if we multiply each of the characteristic functions of the linear oscillator $\psi_1, \psi_2, \cdots, \psi_n, \cdots$ by a corresponding phase factor $e^{i\epsilon_n}$ where ϵ_n is arbitrary, we get a new set of normalized wave functions which is just as good as the first set. This transformation of the characteristic functions multiplies each of the matrix elements in turn by a phase factor of the form $e^{i(\epsilon_n-\epsilon_m)}$. Consequently the hypothesis *d*) holds only for a particular choice of the phase constants of the basic characteristic functions.

Then in view of the hypotheses stated above

$$P^{(x)}(n_1', \cdots, n_f'; n_1'', \cdots, n_f'')e^{2\pi i\nu(n';n'')} \rightarrow P^{(x)}(\tau_1, \cdots, \tau_f)e^{(2\pi i/h)(\sum_k \tau_k \partial E/\partial n_k)t}.$$

Thus Eq. (114) reduces to

$$\overline{P^{(x)}} = \sum_{\tau_1,\cdots,\tau_f} P^{(x)}(\tau_1, \cdots, \tau_f)e^{(2\pi i/h)(\sum_k \tau_k \partial E/\partial n_k)t} \sum_{n_1',\cdots,n_f'} C_{n_1',\cdots,n_f'} C^*_{n_1'-\tau_1,\cdots,n_f'-\tau_f}.$$

Because of the assumed normalization of the wave packet and assumption e) this reduces to the form

$$\overline{P^{(x)}} = \sum_{\tau_1,\cdots,\tau_f} P^{(x)}(\tau_1, \cdots, \tau_f)e^{i(\sum_k \tau_k \epsilon_k)}e^{(2\pi i/h)(\sum_k \tau_k \partial E/\partial n_k)t} \qquad (115)$$

which is similar to that of Eq. (112). Now we have already shown in Part I, Section 2, 4 and 5 that there is an asymptotic agreement between the values of the energy differences predicted by the wave mechanics and by the Bohr theory in the case of a linear oscillator. The same agreement is found in the case of the hydrogen atom problem, and may be proved for the general case of any conditionally periodic motion from the work of Van Vleck.[52] But according to the Bohr theory the values of the classical frequencies $\nu_1, \nu_2, \cdots, \nu_k, \cdots$ are asymptotically equal to the corresponding coefficients $h^{-1}(\partial E/\partial n_k)$ in the exponents of Eq. (115). Thus the frequency factors of the terms in Eq. (115) are the same as the frequency factors in the corresponding terms of Eq. (112), provided only that the wave packet of Eq. (115) is adjusted to the initial conditions characteristic of the classical motion to which Eq. (112) refers. The agreement between the matrix coefficients

$$P^{(x)}(\tau_1, \cdots, \tau_f)e^{i(\sum_k \tau_k \epsilon_k)} = e^{i(\sum_k \tau_k \epsilon_k)} \lim [P^{(x)}(n_1', \cdots, n_f'; n_1'-\tau_1, \cdots, n_f'-\tau_f)]$$

of the one series and the Fourier amplitudes

$$P^{(x)}{}_{\tau_1,\cdots,\tau_f}$$

of the other, can be proved directly by the method used by Van Vleck, or can be inferred from the fact that the motion of the center of gravity of the wave packet in configuration space is identical with the motion of the representative point of the classical system in the same space. (Cf. Hamilton's equations of motion, Section 6, 5).

SECTION 7. TRANSFORMATION THEORY

1. Operators and Canonical Transformation

In the so-called "q-number theory" of Dirac the same non-commutative algebraic machinery was used as in the matrix theory of Born, Heisenberg, and Jordan, but the symbols which the latter authors use for matrices were given a less specific interpretation. Dirac called these symbols "q-numbers" to distinguish them from ordinary numbers or "c-numbers" which obey the

usual commutation rule for algebraic multiplication. These q-numbers were shown by him to be capable of representation by matrices in special cases. The matrix representation, however, is subject to limitations; e.g., in case angles, or "angle variables", are used as coordinates no corresponding matrices satisfying the commutation rules (89) exist. Also if problems involving continuous spectra are to be dealt with, one must use continuous matrices or matrices of partly continuous character, and convergence difficulties cause trouble.[53] These difficulties led Born and Wiener[54] to introduce operators instead of matrices for the discussion of aperiodic motions, and later Schrödinger,[55] Dirac,[56] and Eckart[57] identified the q-numbers with the operators for the corresponding classical quantities as defined in the Schrödinger theory and as used in the preceding sections of this paper (except for the difference in sign noted in the final foot-note of Part I). According to the conventions adopted in this paper the operators associated with the co ordinate q and the conjugate momentum p are*

$$O_q = [q\times] = O_q^* \; ; \; O_p = -(h/2\pi i)\partial/\partial q = -O_p^*. \tag{116}$$

where $[q\times]$ denotes the operation of multiplying by q. The starred symbols represent the operators used by Schrödinger, Eckart, and Dirac. To avoid a confusing inversion of the usual order of multiplication it will be convenient in this section to use these latter operators and to adopt Eckart's symbolism for them. Thus Q is to replace the symbol $[q\times]$ and P is to represent the operator

$$P = O_p^* = (h/2\pi i)\partial/\partial q. \tag{116a}$$

If A and B are any two operators and if $f(q)$ is any function of the q's, then the operators C and D which satisfy the identical relations

$$Cf = Af + Bf \; ; \; Df = A(Bf),$$

are called the *sum* of the operators A and B, and the *product* of A and B respectively. We indicate these relations by the operator equations

$$C = A + B; \quad D = AB.$$

From the second of these equations it is obvious that the operators P and Q satisfy the same commutation rules as the matrices. Thus, since

$$\frac{h}{2\pi i}\left[\frac{\partial}{\partial q_k} q_k\psi - q_k\frac{\partial\psi}{\partial q_k}\right] = \frac{h}{2\pi i}\psi$$

[53] The continuous and mixed types of matrix are sufficiently defined in the fine print on p. 201 of Part I. The difficulties encountered in their use are briefly discussed by Born in his Massachusetts Institute of Technology lectures "Problems of Atomic Dynamics," Cambridge, 1926, at the end of Lecture 19, and by J. v. Neuman in his article "Mathematische Begrundung der Quantenmechanik," Göttingen Nachrichten (1927), p. 1.

[54] M. Born and N. Wiener, Zeits. f. Physik **36,** 174 (1926).

[55] E. Schrödinger, Ann. der Physik **79,** 734 (1926).

[56] P. A. M. Dirac, Proc. Roy. Soc. **A112,** 661 (1926).

[57] C. Eckart, Phys. Rev. **28,** 711 (1926).

* These are the operators to be applied to ψ_q. Cf. p. 44.

for any function $\psi(q)$, we have the identical operator equation

$$P_kQ_k-Q_kP_k=h/2\pi i. \tag{117}$$

The other operator products are commutable and we can summarize the commutation rules by the equations,

$$P_kQ_l-Q_lP_k=(h/2\pi i)\delta_{kl}\ ;\quad P_kP_l-P_lP_k=Q_kQ_l-Q_lQ_k=0. \tag{117a}$$

In developing the general transformation theory of quantum dynamics, Eq. (117) may be taken as the definition of canonical operators in all ordinary cases where the q's assume continuous ranges of values.[57a] The operators P, Q clearly obey the associative law of multiplication and the commutative, distributive, and associative laws of addition.

The reciprocal or inverse operators of Q and P are defined by

$$Q^{-1}=\left[\frac{1}{q}\times\right];\quad P^{-1}=\frac{2\pi i}{h}\int dq \tag{118}$$

in accordance with the rule

$$QQ^{-1}=I=[1\times]\quad PP^{-1}=P^{-1}P=I=[1\times]. \tag{119}$$

With the aid of these conventions we can manipulate the P's and Q's as we do matrices and can deal with any function of the P's and Q's involving only integral powers of the former. The derivative of an operator function $f(X)$ with respect to an argument operator X can be defined in accordance with Eq. (96) by the rule

$$\frac{d}{dX}f(X)=\lim_{A\to 0}\left(\frac{f(X+A)-f(X)}{A}\right) \tag{120}$$

where A is the operator $[a\times]$. When applied to operators P_k, Q_k, which obey the commutation rules, this definition is equivalent to the equations

$$\left.\begin{aligned}\frac{\partial f(P,Q)}{\partial P_k}&=\frac{2\pi i}{h}[fQ_k-Q_kf]=[Q_k,f]\\ \frac{\partial f(P,Q)}{\partial Q_k}&=\frac{2\pi i}{h}[P_kf-fP_k]=[f,P_k]\end{aligned}\right\}. \tag{121}$$

There is a difficulty, however, in connection with the time derivatives needed for the formulation of Hamilton's equations of motion, since the operators P and Q are not functions of the time operator. Furthermore the Hamiltonian operator $H(P, Q)$ is not a function of the time operator, so that so far as the writers can see, the Hamiltonian equations can be applied to the operators P, Q, $H(P, Q)$ only as definitions of the operators $\dot{P}$ and $\dot{Q}$.[58] This

[57a] In case any one of the coordinates q_k assumes only discrete values, e.g. when the energy is used as a coordinate, the operator $(h/2\pi i)\partial/\partial q_k$ has no meaning when applied to $\psi(q)$. Then Eq. (117) fails as do the corresponding matrix relations (89).

[58] Eckart assumes the Hamiltonian equations to be valid for suitable operators and then defines the momentum operators P_k and the time derivatives $\dot{P}_k$ and $\dot{Q}_k$ so as to reduce the equations of motion to identities. Dirac does not discuss the question at all.

means that the equations of motion no longer give any hold on the characteristic value problem.

The operator calculus is useful, however, in dealing with coordinate transformations, and to this problem we now address ourselves. Let the Hamiltonian function of a system be given in Cartesian coordinates, and let us suppose that we wish to make a change of variables. The simplest case is that of an ordinary point transformation in which the old coordinates $x_1, \cdots, x_n$ are functions of the new coordinates $q_1, \cdots, q_n$ and are independent of the momenta. Thus let

$$\begin{aligned} x_1 &= f_1(q_1, \cdots, q_n) \\ x_2 &= f_2(q_1, \cdots, q_n) \\ &\cdots\cdots\cdots \end{aligned} \tag{122}$$

The corresponding operator equations are simply

$$\begin{aligned} X_1 &= f_1(Q_1, \cdots, Q_n) \\ X_2 &= f_2(Q_1, \cdots, Q_n) \\ &\cdots\cdots\cdots \end{aligned} \tag{123}$$

Let the momenta conjugate to the original coordinates be designated by the symbols $p_1^0, \cdots, p_n^0$. The corresponding operators are

$$P_k^0 = (h/2\pi i)\partial/\partial x_k. \tag{124}$$

Similarly we denote the momenta conjugate to the q's by $p_1, \cdots, p_n$ and we *define* the corresponding operators by

$$P_k = (h/2\pi i)\partial/\partial q_k. \tag{125}$$

It is a straightforward job of differentiation to compute the P_k^0's in terms of the P_k's by means of Eqs. (122). The result will be of the form

$$P_k^0 = \sum_l g_{kl}(Q) P_l = \phi(Q_1, \cdots, Q_n ; P_1, \cdots, P_n). \tag{126}$$

By substitution from formulas (123) and (125) we can derive an expression for the original Hamiltonian operator in terms of the P's and Q's. We thus accomplish the change of variables

$$H^0(P^0, X) = H^0[\phi(P,Q), f(Q)] = H^\dagger(P,Q). \tag{127}$$

The process just described is exactly equivalent to an ordinary change of variables in a partial differential equation, so that the solutions of the original equation[59]

$$H^0(P^0, Q^0)\psi^*(x) = -(h/2\pi i)\partial\psi^*(x)/\partial t \tag{128}$$

when expressed in terms of the new coordinates become solutions of the new differential equation

$$H^\dagger(P,Q)\psi^{*\dagger}(q) = -(h/2\pi i)\partial\psi^{*\dagger}(q)/\partial t \tag{129}$$

[59] This is the conjugate of Eq. (1).

with

$$\psi^{*\dagger}(q)=\psi^*[f(q)].$$

This change of variables will in some cases simplify the problem of determining the energy levels and wave functions but it does not reduce the problem in the new coordinates to a form equivalent to that which it had in the original coordinates. This is evident when we note that by changing the variables in the original normalization equation we obtain

$$\int_\infty \psi\psi^* dx_1 \cdots dx_n = \int_\infty \psi^\dagger \psi^{*\dagger} \Delta dq_1 \cdots dq_n = 1 \tag{130}$$

where Δ denotes the Jacobian or functional determinant of the system of functions $f_1, \cdots, f_n$ with respect to the arguments $q_1, \cdots, q_n$; i.e.,

$$\Delta = \frac{\partial(f_1, \cdots, f_n)}{\partial(q_1, \cdots, q_n)}. \tag{131}$$

Thus the solutions of Eq. (128) still have a character which relates them to the original Cartesian coordinate system. On this account we may designate them with a subscript x in accordance with the notation of footnote 51, Part I. The probability that the system has a configuration corresponding to the volume element $dx_1, \cdots, dx_n$ of Cartesian coordinate space is $\psi_x\psi_x^* dx_1 \cdots dx_n$, but the probability that it has a configuration corresponding to the volume element $dq_1 \cdots dq_n$ of the new non-Cartesian coordinate space is $\psi_x^\dagger \psi_x^{*\dagger} \Delta dq_1 \cdots dq_n$. The functions $\psi_x^\dagger$ have only conditional orthogonality in the q-space, and there is a further serious difficulty about computing mean values of the p's which will be indicated below (see especially footnote 60).

To avoid these difficulties we introduce a new wave-function

$$\psi_q = \psi_x^\dagger \Delta^{1/2}. \tag{132}$$

Since

$$\psi_x\psi_x^* dx_1 \cdots dx_n = \psi_q\psi_q^* dq_1 \cdots dq_n, \tag{133}$$

it is evident that ψ_q is the probability amplitude for the coordinate system $q_1, \cdots, q_n$. This equation also shows that the orthogonality of the ψ_x functions in Cartesian coordinate space carries with it the orthogonality of the ψ_q in the q-space. Reference to Sect. 3, Part I will show that we can compute the mean value of the momentum p_k from the formula[60]

$$\overline{p_k} = \int_\infty \psi_q^* \left(-\frac{h}{2\pi i}\frac{\partial}{\partial q_k}\right) \psi_q dq_1 \cdots dq_n = \int_\infty \psi_q P_k \psi_q^* dq_1 \cdots dq_n. \tag{134}$$

[60] $\overline{P}_k$ is defined by an equation analogous to Eq. (93) of Part I. Its evaluation depends on the determination of the corresponding probability amplitude which is to be carried through by Fourier analysis as in Sect. 1, 8 of Part I. In order to carry out this analysis it is clearly essential that we start with ψ_q.

The operator P_k is by Eqs. (125) and (90) clearly self-adjoint with respect to the q coordinate system, and hence the values of $\overline{p_k}$ given by Eq. (134) are necessarily real.

To get the differential equation for this modified wave function we have only to make the substitution of $\psi_q\Delta^{-1/2}$ for $\psi_x^\dagger$ in Eq. (129). We find

$$\Delta^{1/2}H^\dagger(P,Q)\Delta^{-1/2}\psi_q^* = -\left(\frac{h}{2\pi i}\right)\frac{\partial\psi_q^*}{\partial t}. \tag{135}$$

In other words the correct Hamiltonian operator to be used in connection with the coordinate system $q_1, \cdots, q_n$ is

$$H_q(P,Q) = \Delta^{1/2}H^\dagger(P,Q)\Delta^{-1/2}. \tag{136}$$

The matrix elements of H_q with respect to the characteristic functions of Eq. (135) are the same as the matrix elements of H^0 with respect to the corresponding characteristic functions of Eq. (128), and we may conclude that H_q is self-adjoint with respect to the q coordinate system.[61]

Both steps in the derivation of H_q can be regarded as examples of *canonical transformations*. These transformations are re-statements of the initial problem in terms of new variables which, like the original ones, are grouped in canonical pairs. In the general case these transformations involve a mixing of the q's and p's like the contact transformations of classical mechanics. Thus we may define a transformation as the setting up of a new problem by means of a substitution of the form

$$Q^0 \rightarrow F(P,Q)\ ; \quad P^0 \rightarrow G(P,Q). \tag{137}$$

This converts $H^0(P^0, Q^0)$ into $H(P, Q)$. In case Eqs. (137) are of such character that

$$P_kQ_k - Q_kP_k = \frac{h}{2\pi i}$$

if

$$P_k^0Q_k^0 - Q_k^0P_k^0 = \frac{h}{2\pi i},$$

we say that the transformation is canonical. In this case we may interpret P_k as $(h/2\pi i)\partial/\partial q_k$ in accordance with Eq. (125), and then use $H(P, Q)$ to set up a Schrödinger equation for $\psi(q)$.

There are two standard forms of canonical transformation used in wave mechanics. One of these is described by the substitutions

$$\begin{aligned} Q^0 &\rightarrow T^{-1}(P,Q)\ Q\ T(P,Q) \\ P^0 &\rightarrow T^{-1}(P,Q)\ P\ T(P,Q) \end{aligned} \tag{138}$$

in which $T(P, Q)$ is an arbitrary operator function and $T^{-1}(P, Q)$ is its reciprocal. This substitution has the property of carrying over any operator

[61] In carrying through the computation of H_q from H^0 the formulas of Podolsky (Note 51, Part I) or of Schrödinger (Note 50, Part I) can be used to advantage.

function expressible as a power series in the P^0's and Q^0's into a corresponding function of the new variables according to the simple rule

$$F(P^0,Q^0)\rightarrow T^{-1}(P,Q)F(P,Q)T(P,Q). \tag{139}$$

The converse substitution is obviously

$$Q\rightarrow T(P^0,Q^0)Q^0T^{-1}(P^0,Q^0)\ ; \qquad P\rightarrow T(P^0,Q^0)P^0T^{-1}(P^0,Q^0). \tag{140}$$

From the rule of Eq. (139) it follows that the transformation is canonical, and it seems probable that it is in fact the most general type of canonical transformation.[62]

The transformed Hamiltonian function has the form

$$H(P,Q)=T^{-1}(P,Q)H^0(P,Q)T(P,Q) \tag{141}$$

and the new Schrödinger equation obtained from the transformation is

$$T^{-1}(P,Q)H^0(P,Q)T(P,Q)\psi^*(q)=-\frac{h}{2\pi i}\frac{\partial\psi^*(q)}{\partial t}. \tag{142}$$

This is equivalent to

$$H^0(P,Q)T(P,Q)\psi^*(q)=-\frac{h}{2\pi i}\frac{\partial}{\partial t}[T(P,Q)\psi^*(q)] \tag{143}$$

if $T(P, Q)$ is independent of the time. If we set

$$\psi^{*0}(q)=T(P,Q)\psi^*(q)\ ;\ \psi^*(q)=T^{-1}(P,Q)\psi^{*0}(q) \tag{144}$$

Eq. (143) becomes identical with the initial differential equation

$$H^0(P^0,Q^0)\psi^{*0}(q^0)=-(h/2\pi i)\partial\psi^{*0}(q^0)/\partial t \tag{145}$$

except for the change in the symbols for the independent variables. It follows that for every solution of Eq. (145) there is a corresponding solution of Eq. (142) given by Eq. (144). It is assumed by London[63] that if $\psi^0(q^0)$ satisfies the boundary conditions appropriate to Eq. (145), the corresponding solution of Eq. (142) will satisfy the boundary conditions appropriate to the latter equation, but although this hypothesis is verified in various special cases it needs further examination. If we accept this assumption we may conclude that since the characteristic functions of Eq. (145) are derivable from those of (142) the latter equation is in fact a restatement of the original problem in different language.

It is worthy of note that by the use of a suitable exponential form for the operator $T(P, Q)$ it is possible to pass from Cartesian coordinates to spherical coordinates or to angle variables in which the region of definition of the transformed functions $\psi(q)$ is quite different from the region of definition of the initial functions $\psi^0(q^0)$.

[62] Jordan (Zeits. f. Physik **37**, 383 (1926)) has given a proof that the corresponding type of matrix transformation is the most general type of canonical matrix transformation.

[63] F. London, Zeits. f. Physik **40**, 193 (1926).

As a simple example of this type of canonical transformation we may cite the second of the two steps in the process of changing from Cartesian coordinates to generalized coordinates described on p. 44. Here we have $T=\Delta^{-1/2}$ and the substitution is

$$Q\rightarrow\Delta^{1/2}(Q')Q'\Delta^{-1/2}(Q')=Q'$$
$$P\rightarrow\Delta^{1/2}(Q')P'\Delta^{-1/2}(Q'),$$

which leaves the coordinate system unchanged.

A second type of canonical transformation is derived from classical analogy. Let $S(Q, P^0)$ denote an arbitrary function of the operators Q and P^0. Let the variables Q and P be defined in terms of the operators Q^0 and P^0 by means of the equations

$$Q^0=\frac{\partial S(Q,P^0)}{\partial P^0}\,;\quad P=\frac{\partial S(Q,P^0)}{\partial Q}\,. \tag{146}$$

If we solve these equations for Q^0 and P^0 in terms of Q and P and substitute for each of the original variables its equivalent in terms of the new ones, we obtain a new Hamiltonian function. Furthermore the operators P, Q satisfy the commutation rules if they are satisfied by P^0 and Q^0 so that the transformation is canonical. The proof of this last proposition is that by a suitable use of exponentials it is possible to convert the transformation (146) into the form of the substitution (138).[64] By the same method one can show that the Schrödinger equation based on the new Hamiltonian has the same characteristic values as the original equation (128).

The first of the two steps required in changing over from Cartesian to generalized coordinates gives a convenient example of a case in which $S(Q, P^0)$ has a simple form. Using the notation of Eqs. (122) and (123) we set

$$S=\sum_l f_l(Q_1, \cdots, Q_n)P_l^0. \tag{147}$$

Then

$$X_k=\frac{\partial S}{\partial P_k{}^0}=f_k(Q_1, \cdots, Q_n) \tag{148}$$

in accordance with Eq. (123). Similarly

$$P_k=\frac{\partial S}{\partial Q_k}=\sum_l \frac{\partial f_l}{\partial Q_k}P_l{}^0=\frac{h}{2\pi i}\sum_l \frac{\partial x_l}{\partial q_k}\frac{\partial}{\partial x_l}\,. \tag{149}$$

This is the inverse of Eq. (126). It shows that P_k is computed by a rule equivalent to the partial differentiation in the discussion on p. 43 and thus insures that the canonical transformation derived from Eq. (147) is an ordinary change of variables.

[64] Cf. P. Jordan, Zeits. f. Physik **38**, 513 (1926). A. Landé, Handbuch der Physik XX, Kap. 8, p. 434. P. A. M. Dirac, Proc. Roy. Soc. **A113**, 621 (1927).

In Article 5 of Sect. 6 we have shown that the classical Hamiltonian equations can be derived from the Hamiltonian equations of the matrix theory in the limiting case of a wave packet whose configuration is defined with precision. This proof requires extension due to the fact that it does not rigorously cover problems in which there is a continuous spectrum of E values but more especially because matrices for a coordinate q_k and its conjugate momentum p_k satisfying the commutation rules do not exist if either one is "quantized" so that it takes on only discrete values.[65] This situation arises when one uses the action and angle variables so important in classical dynamics and in the Bohr theory.

It would be natural to seek a means of bridging the above gap with the aid of operator equations of motion paralleling the matrix equations of motion. This method gets nowhere, however, on account of the impossibility of differentiating the P, Q operators with respect to the time. The following procedure, however, leads to the goal.

Let a system of generalized coordinates $q_1, \cdots, q_n$ be given and let ψ denote the probability amplitude ψ_q which goes with the system. The correct Hamiltonian function H is then self-adjoint with respect to the coordinates $q_1, \cdots, q_n$. Let θ denote an angular coordinate such that ψ is periodic in θ with period 2π. Consider now the time derivative of the mean value of $e^{i\theta}$

$$\frac{d}{dt}\int_\infty \psi^* e^{i\theta}\psi\, dq_1 \cdots dq_n = \int_\infty [\dot{\psi}^* e^{i\theta}\psi + \psi^* e^{i\theta}\dot{\psi}]\, d\tau$$

$$= \frac{2\pi i}{h}\int_\infty \left[\psi^* e^{i\theta} H\left(-\frac{h}{2\pi i}\frac{\partial}{\partial q}, Q\right)\psi - \psi e^{i\theta} H\left(\frac{h}{2\pi i}\frac{\partial}{\partial q}, Q\right)\psi^*\right] d\tau. \qquad (150)$$

Denoting the second of these Hamiltonian operators by $H(P, Q)$ and making use of the fact that it is self-adjoint we obtain

$$\frac{d}{dt}\overline{e^{i\theta}} = \frac{2\pi i}{h}\int_\infty \psi[H(P,Q)e^{i\theta} - e^{i\theta}H(P,Q)]\psi^* dq_1 \cdots dq_n. \qquad (151)$$

This expression can be thrown into a more convenient form by means of the theorem[66]

[65] Cf. P. Jordan, Zeits. f. Physik **44**, 1 (1927). The reader will readily convince himself by actual multiplication that in the hydrogen atom problem the matrices for the azimuthal angle ϕ and its conjugate momentum M_z as defined by Eq. (68) do not satisfy the commutation rule of Eq.(85). The failure of what seems at first glance to be a clean-cut general theorem is to be referred to the restrictions on the preliminary theorem of p. 8.

[66] To prove the theorem note first that

$$P_\theta e^{i\theta} = \frac{h}{2\pi i}\frac{\partial}{\partial\theta}e^{i\theta} = e^{i\theta}\left[\frac{h}{2\pi i}\frac{\partial}{\partial\theta} + \frac{h}{2\pi}\right].$$

Obviously the theorem is true for $F(P_\theta, Q) = F(Q)$ and one can prove that it is true for the sum and product of two functions if it is true for the functions taken separately. The extension to sufficiently general cases follows by induction as in footnote 45. Cf. also P. A. M. Dirac, Proc. Roy. Soc. **A111**, 281 (1926).

$$F(P_\theta, Q)\, e^{i\theta} = e^{i\theta} F\left(P_\theta + \frac{h}{2\pi},\ Q\right). \tag{152}$$

Identifying $F(P_\theta, Q)$ with $H(P, Q)$ we obtain

$$\frac{d}{dt}\overline{e^{i\theta}} = i \int_\infty \psi e^{i\theta} \left[\frac{H\left(P_\theta + \frac{h}{2\pi},\ Q\right) - H(P_\theta, Q)}{h/2\pi} \right] \psi^* dq_1 \cdots dq_n. \tag{153}$$

As a final step one can define the mean velocity associated with the angle θ by the relation

$$\overline{\dot{\theta} e^{i\theta}} = \frac{d}{dt}\overline{e^{i\theta}}.$$

Then passing to the limit of a sharply defined wave packet, the right hand member of Eq. (153) becomes equal to the product of the partial derivative of the classical Hamiltonian into the function $\overline{e^{i\theta}}$. Hence

$$\overline{\dot{\theta}} = \frac{\partial H(\overline{p}, \overline{q})}{\partial \overline{p_\theta}}. \tag{154}$$

This takes care of the difficult type of Hamiltonian equation for a coordinate of the angle-variable type. The other Hamiltonian equations can be derived in a similar manner by the transformation of the time derivatives of p_k and q_k.

2. The Dirac-Jordan Transformation Theory

Various gaps in the preceding discussion are bridged and loose ends tied together by the Dirac-Jordan transformation theory.[67,68] Unfortunately our exposition of their point of view must be very much abridged. In the main we shall follow the fuller and more general treatment of Jordan who has cast the essence of the matter in the form of a few postulates.

As the earlier sections of this paper testify, the quantum mechanics has much to say about "probability amplitudes." These are functions of coordinates, or systems of coordinates, whose squared absolute values determine the absolute or relative probability of the corresponding configurations. The probability depends, of course, on the circumstances and the circumstances which go with such functions are frequently labeled in the symbol for the function. For example, in the case of a one-dimensional oscillator, each of the characteristic functions $\psi_n(x, t)$ is a probability amplitude for the positional coordinate x when the energy E is known to have the value E_n. The "circumstances" are indicated by the subscript n. In this case the independent variable x has a continuous range of values, and accord-

[67] P. A. M. Dirac, Proc. Roy. Soc. **A113**, 621 (1927).

[68] P. Jordan, Zeits. f. Physik **40**, 809 (1927); **44**, 1 (1927). These two papers will be referred to as Jordan I and Jordan II respectively.

ingly $|\psi_n(x, t)|^2 dx$ measures the absolute or relative probability[69] that x has some value in the range from x to $x+dx$. In other cases the circumstances are not to be described so simply. For example, the solutions of the Schrödinger equation in the form which contains the time derivative (I, 38) depend on the initial conditions of the problem, which are fixed by the way in which each system of the ensemble is prepared. Any mode of preparation of the system may be regarded, however, as a specification of certain generalized coordinates,[70] and, following Jordan, one may formally attach subscripts for these coordinates to the symbols for the probability amplitudes. Thus in the case of the one-dimensional problem one can label the wave function $\psi(x, t)$ as $\psi_\beta(x, t)$ to indicate that the initial conditions of the problem are the equivalent of the specification of a suitably chosen "coordinate" β.

In case a coordinate β takes on a discrete series of values instead of a continuous range, there is a finite probability for each possible value of β. The probability of this value is then given directly by the product of the probability amplitude into its conjugate without multiplying in a factor $d\beta$. As an example, consider the probability amplitude for the energy in the case of a one-dimensional oscillator. This is the system of coefficients c_n, or $c(E_n)$, obtained when one analyses $\psi(x, t)$ into a series of characteristic functions $\psi_n(x, t)$. The probability of each discrete energy value is $c_n c_n^*$ while the probability of any energy range in the continuous spectrum is $c(E)c^*(E)dE$.[71]

As examples of probability amplitudes in problems involving more dimensions we cite the characteristic functions of the hydrogen atom problem obtained by separating the variables in spherical coordinates. These give the probabilities of the positional coordinates when the component of angular momentum along the z-axis, the square of the total angular momentum, and the total energy are given predetermined values. The functions $\psi(t,x, y,z)$ and $Q(t, p_x, p_y, p_z)$ met with in Sect. 1, 8 of Part I are probability amplitudes for position and momentum respectively for definite but unspecified modes of preparing the system.

Jordan adopts the symbol $\Phi_{\alpha,\beta}(\beta', q')$ for the probability amplitude of the coordinate system $\beta(=\beta_1, \cdots, \beta_f)$ with respect to the coordinate system $q(=q_1 \cdots, q_f)$; that is, $\Phi_{\alpha,\beta}(\beta', q')$ is the probability amplitude for the coordinates $\beta_1, \cdots, \beta_f$ when the coordinates $q_1, \cdots q_f$ are given specified values. Here the argument symbols are primed in order to distinguish the numerical values β_k', q_k' which the coordinates take on, from the symbols β_k, q_k for the coordinates themselves. The subscript α stands for the set of momenta conjugate to the coordinates β, while p stands for the momenta conjugate to the q's.[72]

[69] The former in case $\psi_n(x, t)$ is normalized.

[70] Here "generalized coordinate" means a function of the Cartesian coordinates and momenta.

[71] Cf. Eq. (I, 85) and the last paragraph on p. 214 of Part I.

[72] The momenta are not uniquely determined by the corresponding coordinates and the probability amplitude varies from one momentum system to another although, as Jordan proves, $|\Phi_{\alpha,\beta}(\beta', q')|^2$ depends only on the β's and q's. Cf. Jordan II, p. 20.

In normalizing the probability amplitudes we have to sum over all discrete values of the arguments and integrate over all continuous ranges. This process we indicate by the symbol $\overline{\sum}$ which degenerates into a simple summation sign or simple integration sign in special cases. As an example of the use of the symbolism we may take Eq. (I, 90) which gives the expansion of an arbitrary function of x, y, z in terms of the space factors of the hydrogen atom problem. The left hand member is a possible probability amplitude for the positional coordinates and the c's constitute the corresponding probability amplitude for the energies, so that the equation can be written in the form

$$\Phi_{p_x,p_q}(x,q) = \overline{\sum_E} \Phi_{p_x,p_E}(x,E)\,\Phi_{p_E,p_q}(E,q) \tag{155}$$

where x stands for the x, y, z coordinate system, q stands for an unspecified set of coordinates whose values fix the function $f(x, y, z)$, and E stands for the coordinate system E, l, m which defines the set of orthogonal functions employed in the expansion. (The momentum conjugate to the energy itself may be identified with the time, but in this case we use the symbol p_E since we have used E for a system of coordinates and not for one coordinate only.) The phrase "system of coordinates" has several different meanings in the quantum mechanics. The simplest type of coordinate system is a set of f parameters $q_1, \cdots, q_f$ which are independent functions of the Cartesian positional coordinates $x_1, \cdots, x_f$. Canonical transformations, however, lead to the introduction of coordinates which are independent functions of both the Cartesian coordinates and their momenta. The transformation theory requires a further modification of this conception in that it treats the values of $q_1, \cdots, q_f$ at each instant of time as forming different coordinate systems. Thus the determination of the coordinates at $t=t_1$ from those at $t=t_0$ is regarded as a coordinate transformation.

In carrying out the normalization we need in some cases an imaginary and magical function $\delta(x)$ first introduced by Dirac.[67] This is defined as a function which vanishes when x is not equal to 0 and whose integral over a range including the point $x=0$ is equal to unity. Strictly speaking, no such function exists, but we use the symbol as a convention to avoid circumlocution. The justification lies in the Dirichlet integral formula

$$\lim_{a\to\infty} \int_{-\infty}^{+\infty} dx F(x) \int_{-a}^{+a} dy\; e^{2\pi i x y} = F(0) \tag{156}$$

If we define $\delta_a(x)$ by the formula

$$\delta_a(x) = \int_{-a}^{+a} dy\; e^{2\pi i x y} \tag{157}$$

we see that when a is very large the properties of $\delta_a(x)$ approach those of the imaginary function $\delta(x)$ although we may not actually allow a to become infinite due to the non-convergence of the integral. The existence of the func-

tion $\delta_a(x)$ makes it possible to carry through many calculations formally just as if there were a genuine function $\delta(x)$, without making errors.[73]

For use in many-dimensional problems, Jordan has extended this conception by introducing a function $\Delta(\beta'-\beta'')$ of the pair of points $\beta_1', \beta_2', \cdots, \beta_f'$ and $\beta_1'', \cdots, \beta_f''$ in β-space. This function is defined to have the properties

$$\Delta(\beta'-\beta'')=0 \quad \text{unless} \quad \beta_1'=\beta_1''; \ \beta_2'=\beta_2''; \cdots \beta_f'=\beta_f''.$$

$$\overline{\sum_{\beta'}}\Delta(\beta'-\beta'')=1. \tag{158}$$

In one dimension it reduces to the Weierstrassian symbol[74] $\delta_{\beta',\beta''}$ if the argument β takes on discrete values, and to the δ function of the preceding paragraph with $x=\beta'-\beta''$ if β assumes continuous values.

By the aid of this symbol one can state the normalization and orthogonality relations for probability amplitudes with both discrete and continuous values of the arguments in the single form

$$\overline{\sum_{q'}}\Phi_{\alpha,p}(\beta',q')\Phi^*_{\alpha,p}(\beta'',q')=\Delta(\beta'-\beta''). \tag{159}$$

With these preliminary ideas in mind, we proceed to state Jordan's fundamental assumptions in the following manner:

(I) *In the case of any two systems of coordinates β, q it is possible to define conjugate systems of momenta α, p such that there exists a probability amplitude*[75]

$$\Phi_{\alpha,p}(\beta',q')$$

having the property that if we perform the operation $\overline{\sum}_{\beta'}$ on the product $\Phi_{\alpha p}(\beta', q')\ \Phi_{\alpha p}{}^*(\beta', q')$ over any region in β-space we obtain the relative or absolute probability of the corresponding configurations of the system (β) for the given values of the coordinates q.

(II) *If we have the probability amplitudes for two coordinate systems with respect to a third we can determine the probability of the two with respect to one another by the rule*

$$\Phi_{\alpha,P}(\beta',O')=\overline{\sum_{q'}}\Phi_{\alpha,p}(\beta',q')\Phi_{p,P}(q',Q'). \tag{160}$$

This rule is exemplified in Eq. (155) above and in Eqs. (42), (43), and (87) of Part I.

(III) *The probability amplitude of a system of coordinates β with respect to the conjugate momenta α is*[76]

[73] Cf. E. H. Kennard, Zeits. f. Physik **44**, 326 (1927).

[74] Cf. Sect. 2, 7 of Part I.

[75] The probability amplitude connected with such a pair of coordinate systems must, of course, be independent of the particular problem under discussion; i.e., it must depend only on the relations existing between the two systems of coordinates.

[76] This rule leads to the identification of $\Phi_{\alpha p}(\beta', q')$ with the wave function which we would have indicated by the symbol $\psi^*(\beta)$ in our old notation.

$$\Phi_{\alpha,-\beta}(\beta',\alpha') = \text{const} \times \exp\left[\left(\frac{2\pi i}{h}\right) \sum_{k=1}^{f} \beta_k' \alpha_k'\right]. \tag{161}$$

This assumption re-states for the more general case the conclusion reached in Sect. 1 for an individual particle, that the momenta corresponding to a Cartesian coordinate system play the part of wave-number components for the waves in coordinate space. It is almost, but not quite, equivalent to the assumption of Eq. (I, 123), since it is applicable in case either β or α takes on a discrete range of values, while Eq. (I, 123) loses its meaning if the coordinate q has only discrete values. It can be regarded as a convenient definition of the relation existing between a coordinate and its canonically conjugate momentum. The subscripts on the symbol $\Phi_{\alpha,-\beta}(\beta', \alpha')$ indicate that if the α's are momenta conjugate to the β's, conversely if we treat the α's as coordinates then the β's with their signs changed become the new momenta. This reciprocal relationship between coordinates and conjugate momenta is a characteristic of the classical theory.

(IV) *If $\Phi_{\alpha p}(\beta', q')$ is the probability amplitude for the coordinate system β with respect to the system q, the corresponding amplitudes for the system q with respect to the system β is $\Phi^*_{\alpha p}(\beta', q')$. Thus*

$$\Phi_{p,\alpha}(q',\beta') = \Phi^*_{\alpha,p}(\beta',q'). \tag{162}$$

For the case in which the q's are canonically conjugate to the β's this rule is merely a corollary on (III).

(V) *If two canonical systems of coordinates β, α and $\bar{\beta}$, $\bar{\alpha}$ have a common coordinate $\beta_1 = \bar{\beta}_1$, the amplitude $\Phi_{\alpha\bar{\alpha}}(\beta', \bar{\beta}')$ contains the factor $\Delta(\beta_1' - \bar{\beta}_1')$. Thus*

$$\Phi_{\alpha,\bar{\alpha}}(\beta',\bar{\beta}') = \Delta(\beta_1' - \bar{\beta}_1')\psi(\beta',\bar{\beta}') \text{ if } \beta_1 = \bar{\beta}_1. \tag{163}$$

This axiom is probably derivable from the others.

These probability amplitudes have certain of the properties of matrices, viz., they are functions of two sets of variables which are equal in number, and furthermore the mode of computing the value of any of these functions for a particular set of values of the variables from related probability amplitudes (Cf. Eq. (160)) is formally similar to the rule for the multiplication of matrices (Cf. Eq. (69)).[77] On this account Jordan introduces the symbol $\Phi_{\alpha p}{}^{\beta q}$ for the whole system of values $\Phi_{\alpha p}(\beta', q')$ and using matrix notation replaces Eq. (160) by

$$\Phi_{\alpha P}{}^{\beta Q} = \Phi_{\alpha p}{}^{\beta q} \Phi_{pP}{}^{qQ}. \tag{165}$$

Eq. (165) is related to (160) just as the ordinary matrix equation

$$\boldsymbol{C} = \boldsymbol{AB}$$

is related to the equation

$$C(m,n) = \sum_k A(m,k)B(k,n).$$

[77] The probability amplitudes differ, however, from the matrices of Sect. 6 in that their rows and columns refer in general to different things (β's and q's respectively) instead of to the same set of variables. Jordan calls them "matrices of the first kind."

Equation (162) now takes the form

$$\Phi_{p\alpha}{}^{q\beta} = \Phi_{\alpha p}{}^{\beta q *}. \tag{166}$$

From Eq. (165) we deduce that

$$\Phi_{\alpha\alpha}{}^{\beta\beta}\Phi_{\alpha p}{}^{\beta q} = \Phi_{\alpha p}{}^{\beta q} \tag{167}$$

which means that as $\Phi_{\alpha p}{}^{\beta q}$ is arbitrary, $\Phi_{\alpha\alpha}{}^{\beta\beta}$ must play the role of an unit matrix. Explicitly written out this becomes

$$\Phi_{\alpha,\alpha}(\beta',\beta'') = \Delta(\beta'-\beta'').$$

The probability that $\beta=\beta'$, when β is by hypothesis equal to β'', is zero unless $\beta'=\beta''$. Furthermore it follows that

$$\Phi_{\alpha p}{}^{\beta q}\Phi_{p\alpha}{}^{q\beta} = \Phi_{p\alpha}{}^{q\beta}\Phi_{\alpha p}{}^{\beta q} = 1. \tag{168}$$

This proves the orthogonality and the completeness[78] of the set of probability amplitudes, for explicitly written out it yields

$$\overline{\sum_{q'}}\Phi_{\alpha,p}(\beta',q')\Phi_{\alpha,p}{}^{*}(\beta'',q') = \Delta(\beta'-\beta''), \tag{169}$$

and

$$\overline{\sum_{\beta'}}\Phi_{\alpha,p}(\beta',q')\Phi_{\alpha,p}{}^{*}(\beta',q'') = \Delta(q'-q''). \tag{170}$$

In addition to the probability amplitudes, the development of the theory requires the introduction of linear operators for the various dynamical variables. The differential operators used hitherto are unsatisfactory, however, since they cannot be applied to functions in which the independent variables take on discrete values. Hence Jordan invents integral operators to take their place.

The fundamental characteristic of the operators previously employed is indicated by the much-used equation

$$O_\alpha\left(q,\frac{\partial}{\partial q}\right)\psi_\alpha(q) = \alpha\psi_\alpha(q) \tag{171}$$

whose solution is to give a probability amplitude for the q's corresponding to a fixed value of α. In terms of our present notation we may require of the operator for a dynamical variable β_1 that it shall have the property

$$O_{\beta_1}\Phi_{p\alpha}(q',\beta'') = \beta_1''\Phi_{p\alpha}(q',\beta''). \tag{172}$$

An equally satisfactory operator is one which transforms a function of q'' into a function of q' according to the rule

$$O_{\beta_1}\Phi_{p\alpha}(q'',\beta'') = \beta_1''\Phi_{p\alpha}(q',\beta''). \tag{173}$$

[78] Cf. Courant-Hilbert, Chap. II, Article 1, 3 Eq. (9).

An integral operator with this characteristic is[79]

$$O_{\beta_1} \equiv {}_{q'',}\beta_{1_{q'}} = \overline{\sum_{\beta'}}\ \overline{\sum_{q''}} \Phi_{\alpha p}{}^{*}(\beta', q')\beta_1' \Phi_{\alpha p}(\beta', q'') \cdots \tag{174}$$

as the reader can easily verify. (The notation ${}_{q'',}\beta_{1_{q'}}$ is used to indicate that the operator carries a function of q'' into a function of q'.)

In special cases this operator reduces to the corresponding operator of Sect. 7, 1. Suppose, for example, that we identify β_1 in Eq. (174) with q_1. Then we obtain the operator for q_1 with respect to the probability amplitudes of the system of coordinates to which q_1 belongs. Eq. (174) reduces to

$$\left.\begin{aligned} {}_{q'',}q_{1_{q'}} &= \overline{\sum_{q'''}}\ \overline{\sum_{q''}} \Phi_{pp}{}^{*}(q''', q')q_1''' \Phi_{pp}(q''', q'') \cdots \\ &= \overline{\sum_{q'''}}\ \overline{\sum_{q''}} \Delta(q'''-q')q_1''' \Delta(q'''-q'') \cdots \end{aligned}\right\} \tag{175}$$

When applied to any probability amplitude with argument q'', say $\Phi_{p\alpha}(q'', \beta')$ this changes q'' into q' and multiplies by q'. Hence it is equivalent to the operator Q of Sect. 7, 1.

Consider next the operator for one of the momenta p_1 with respect to the same coordinates

$$\begin{aligned} {}_{q'',}p_{1_{q'}} &= \overline{\sum_{p'}}\ \overline{\sum_{q''}} \overset{*}{\Phi}_{-q,p}(p', q')p_1' \Phi_{-qp}(p', q'') \cdots \\ &= \overline{\sum_{p'}}\ \overline{\sum_{q''}} e^{(2\pi i/h)p_1'(q_1'-q_1'')} p_1' \Delta(q_2'-q_2'') \cdots \Delta(q_f'-q_f'') \cdots . \end{aligned} \tag{176}$$

Evidently the operator merely substitutes q_k' for q_k'' for every value of the subscript k except $k=1$. In case q_1 and p_1 both vary continuously it can be proved from Dirichlet's integral that

$${}_{q'',}p_{1q'} \Phi_{p,\alpha}(q'', \beta') = \text{const.} \times \left(\frac{h}{2\pi i}\right) \left[\frac{\partial \Phi_{p,\alpha}(q'', \beta')}{\partial q_1''}\right]_{q''=q'} . \tag{177}$$

In this case also we see that the integral operator is equivalent to the corresponding differential operator P of Sect. 7, 1.

Consider next the matrices of the type used in Sect. 6 formed from the operator ${}_{q'',}\beta_{1_{q'}}$ and the probability amplitudes $\Phi_{p,\tau}(q'', \mu'')$ and $\Phi_{p,\tau}(q', \mu')$. In analogy with Eq. (68) we define the element (μ', μ'') of this matrix as*

$$\tilde{B}_1(\mu', \mu'') = \overline{\sum_{q'}} \Phi_{p,\tau}{}^{*}(q', \mu')_{q'',}\beta_{1q'} \Phi_{p,\tau}(q'', \mu'') . \tag{178}$$

[79] We here define the operator by means of the probability amplitude where in other cases we have defined the probability amplitude in terms of the operator. This procedure is a consequence of the fact that probability amplitudes for canonically conjugate quantities are treated as fundamental in the present theory.

* The use of the wavy line in $\tilde{B}_1(\mu', \mu'')$ does not refer to a transposition of the indices μ' and μ'' as in Sect. 6, 1 but is used merely to distinguish the matrices of Eqs. (178) and (179) from those of (180).

Hence

$$\tilde{B}_1(\mu',\mu'') = \overline{\sum_{\beta'}} \Phi_{\alpha\tau}{}^*(\beta',\mu')\beta_1'\Phi_{\alpha\tau}(\beta',\mu'') . \tag{179}$$

The second of the above expressions for $\tilde{B}_1(\mu', \mu'')$ becomes the analog of Eq. (I, 93) if we identify μ' and μ''. Thus this very general type of matrix formula contains within itself both the matrix elements of Sect. 6 and the mean value formulas of Sect. 3.

Finally, if we form the matrix $B_1(q', q'')$ of ${}_{q''}\beta_{1_{q'}}$ with respect to the coordinate system q, p we obtain

$$B_1(q',q'') = \overline{\sum_{\beta'}} \Phi_{\alpha p}{}^*(\beta',q')\beta_1'\Phi_{\alpha p}(\beta',q'') . \tag{180}$$

Comparing with Eq. (174) we see that

$${}_{q''}\beta_{1q'} = \overline{\sum_{q''}} B_1(q',q'') \cdots . \tag{181}$$

Thus the operator of β_1 with respect to the coordinate system q, p is equivalent to a linear transformation whose matrix is $B_1(q', q'')$.

Matrices of the type $\tilde{B}_1(\mu', \mu'')$ and $B_1(q', q'')$ are called "*matrices of the second kind.*" If we introduce the matrix

$$\beta_1{}^{\beta\beta} = \beta_1'\Delta(\beta'-\beta'') \tag{182}$$

we can express Eq. (180) in the form

$$B_1 = (\Phi_{\alpha p}{}^{\beta q})^{-1}\beta_1{}^{\beta\beta}(\Phi_{\alpha p}{}^{\beta q}) . \tag{183}$$

We call $\beta_1{}^{\beta\beta}$ the matrix of β_1 with respect to a system of coordinates of which it is itself a member, while B_1 is the matrix of β_1 with respect to the coordinate system q, p. This means, for example, that the probability amplitude $\psi_{nlm}(x,' y, z)$ in the hydrogen atom problem can be interpreted as the matrix of a transformation which carries the matrix of x (or y or z) with respect to an x, y, z coordinate system over into the matrix of x with respect to the "coordinate system" n, l, m.[80] Thus (for $\beta_1 = x$) Eq. (183) goes over into the form

$$X(n',l',m';n'',l'',m'') = \int_\infty \psi^*_{n''l''m''}(x,y,z)\, x\, \psi_{n'l'm'}(x,y,z)\, dx\, dy\, dz . \tag{184}$$

If the matrix B_1 in Eq. (183) is known, we can use the equation to determine the probability amplitude $\Phi_{\alpha p}{}^{\beta q}$. Multiplying both sides in front by $\Phi_{\alpha p}{}^{\beta q}$ and writing out the explicit expression for the matrix element we obtain

[80] This use of the quantum numbers as a "coordinate system" is due to Dirac (ref. 67). More correctly we should say that the coordinate system consists of the set of dynamical variables determined by the quantum numbers; e.g., the total angular momentum, etc.

$$\overline{\sum_{q''}}\Phi_{\alpha p}(\beta', q'')B_k(q'', q') = \overline{\sum_{\beta''}}\beta_k'\Delta(\beta'-\beta'')\Phi_{\alpha p}(\beta'', q') = \beta_k'\Phi_{\alpha p}(\beta', q'). \tag{185}$$

If the coordinates q have discrete spectra only, Eq. (185) is a homogeneous linear equation with the elements of the matrix $\Phi_{\alpha p}{}^{\beta q}$ as unknowns. If we allow q'' to take on all possible values we obtain an infinite set of simultaneous linear equations of which Eq. (I, 135) may be regarded as a special case where β is identical with the energy.

On the other hand, the above equation can be reduced, under suitable circumstances, to differential form. As B_k is Hermitian,

$$\overline{\sum_{q''}}B_k(q', q'')\Phi_{p\alpha}(q'', \beta') = \beta_k'\Phi_{p\alpha}(q', \beta'). \tag{186}$$

Define ${}_{q'}B_{kq'}$ as the operator

$$\overline{\sum_{q''}}\Delta(q'-q'')\ \overline{\sum_{q'}}B_k(q'', q')\cdots.$$

${}_{q'}B_{kq'}$ is equivalent to ${}_{q''}\beta_{kq'}$ or to $\overline{\sum_{q''}}B_k(q', q'')$ except that it converts a function of q' into a function of q'. This reduces Eq. (186) to the form

$$\left\{{}_{q'}B_{kq'}-\beta_k'\right\}\Phi_{p\alpha}(q', \beta') = 0. \tag{187}$$

Finally, if the q coordinate system is continuous, so that

$${}_{q'}p_{kq'} = \frac{h}{2\pi i}\frac{\partial}{\partial q'};\quad {}_{q'}q_{kq'} = [q_k'\times]$$

and if ${}_{q'}B_{kq'} = g_k({}_{q'}q_{kq'};\ {}_{q'}p_{kq'})$, we have

$$\left\{g_k\left(q', \frac{h}{2\pi i}\frac{\partial}{\partial q'}\right)-\beta_k'\right\}\Phi_{p\alpha}(q', \beta') = 0. \tag{188}$$

This is paralleled by

$$\left\{f_k\left(q', \frac{h}{2\pi i}\frac{\partial}{\partial q'}\right)+\frac{h}{2\pi i}\frac{\partial}{\partial\beta_k'}\right\}\Phi_{p\alpha}(q', \beta') = 0 \tag{188'}$$

in case β_k has a continuous range of values. Here, $f_k(q,\ p) = \alpha_k$. Suitable specialization of Eq. (188′) leads to the differential equation for Q referred to on p. 173 of Part I. If β_k is the energy and $g_k(q,\ p)$ is the Hamiltonian, (188) is the Schrödinger equation without the time. In general it is identical with Eqs. (I, 124) and (171).

The wave functions in the second Shcrödinger equation are probability amplitudes associated with the transformation from a fixed coordinate system ξ to a variable coordinate system q_t. The problem of determining $\psi(q,\ t)$

from the initial conditions as defined by $\psi(q, 0)$ is the problem of passing from $\Phi(q_0, \xi)$ to $\Phi(q_t, \xi)$, which is solved by use of the formula

$$\Phi(q_t, \xi) = \sum\!\!\!\!\!\!\int_{q_0} \Phi(q_t, q_0)\Phi(q_0, \xi)$$

if $\Phi(q_t, q_0)$ can be found. Kennard[73] has discussed the problem in detail and has proved that if $\Phi(q_t, \xi)$ is formed in this way it is then a solution of an equation which differs from Eq. (I, 38) only through a term which is of no physical importance.

Eqs. (185), (186), and (187) can be paralleled by corresponding equations in which the matrix of β_k (*i.e.*, B_k) is replaced by the matrix of the momentum α_k conjugate to β_k. At the same time we have to replace the factor $\beta_{k'}$ on the right by ${}_{\beta'}\alpha_{k_{\beta'}}$ [Cf. Eq. (178)]. Then we have, for example

$$\{ {}_{q'}A_{k_{q'}} - {}_{\beta'}\alpha_{k_{\beta'}}{}^* \} \Phi_{p\alpha}(q', \beta') = 0. \tag{187a}$$

This equation, regarded as a functional equation for the probability amplitude $\Phi_{p\alpha}(q', \beta')$ has the same solutions as Eq. (187). The equivalence of the solutions of the two equations is an expression of the fact that β_k and α_k are canonically conjugate quantities. If we now, starting with an arbitrary set of canonically conjugate coordinates and momenta (q, p) having continuous spectra, form arbitrary operator functions[81]

$$\left.\begin{aligned} {}_{q'}B_{k_{q'}} &= g_k({}_{q'}q_{k_{q'}}\,;{}_{q'}p_{k_{q'}}) \\ {}_{q'}A_{k_{q'}} &= f_k({}_{q'}q_{k_{q'}}\,;{}_{q'}p_{k_{q'}}), \end{aligned}\right\} k = 1, \cdots, f \tag{189}$$

we can set up a corresponding pair of equations of the form of Eq. (187) and (187a).[82] If the spectra for the β's and α's can be defined such that these equations have a simultaneous solution, then the coordinate system (β, α) thus defined is canonical. This rather formidable method of testing a set of proposed transformation equations like (189) reduces to the one previously used if the β spectrum, like the p and q spectra is continuous. Then (187a) goes over into

$$\left\{ f_k\left(q', \frac{h}{2\pi i}\frac{\partial}{\partial q'}\right) + \frac{h}{2\pi i}\frac{\partial}{\partial \beta_k'} \right\} \Phi_{p\alpha}(q', \beta') = 0. \tag{188a}$$

and the condition for the identity of the solutions of (188) and (188a) becomes[83]

$$f_k g_k - g_k f_k \equiv 1. \tag{190}$$

A canonical transformation can also be set up by means of the matrix equations

$$\bar{q}_k = (\Phi_{p\alpha}{}^{q\beta})^{-1} q_k (\Phi_{p\alpha}{}^{q\beta})$$
$$\bar{p}_k = (\Phi_{p\alpha}{}^{q\beta})^{-1} p_k (\Phi_{p\alpha}{}^{q\beta})$$

[81] Cf. Eqs. (137).

[82] The operator ${}_{\beta'}\alpha_k{}^*{}_{\beta'}$ is defined by Eq. (176).

[83] Cf. Jordan I, §5.

where $\Phi_{p\alpha}{}^{q\beta}$ is an arbitrary function of the q's and β's satisfying the conditions (166) and (168). If the transformation is to preserve the Hermitian character of the Hamiltonian matrix (which is equivalent to keeping the Hamiltonian operator self-adjoint), it is necessary that $\Phi_{p\alpha}{}^{q\beta}$ be independent of an interchange in the numerical values of the q's and β's.

For further details and applications of this beautiful theory the reader is referred to the original papers of Dirac and Jordan.[84]

ERRATA

EDWIN C. KEMBLE

Physical Review Supplement 2, October, 1929

Page 174, last line. The right hand member of the equation should be multiplied by dx.

Page 192, Eq. (97) should read

$$\iiint_{\infty}(\psi_1\psi_2^*+\psi_2\psi_1^*)dx\,dy\,dz=\sum\iiint_{\infty}(Q_1Q_2^*+Q_2Q_1^*)dp_x dp_y dp_z.$$

Eq. (98) is best derived by means of Eq. (91) of Part II.

Page 193, Eq. (102). The sign of the right hand member should be reversed.

Page 193, line 13 from bottom. Read (94) and (96) for (94) and (95)

Page 203, Eq. (137) should read $\sum_k |c_{kn}|^2=1$.

[84] Cf. also D. Hilbert, L. Nordheim, and J. V. Neuman, Math. Ann. 98, 1 (1927).

THE INTERPRETATION OF BAND SPECTRA. PARTS I, IIa, IIb.

By Robert S. Mulliken
University of Chicago

SYMBOLS AND NOTATION†

The following table is an index and summary of the symbols and notation used in Part I of this review. Some of the symbols used here are different from those customary in papers on band spectra, but are in accordance with an agreement recently arrived at by band spectroscopists, and soon to be published in the Physical Review.* Such changes have been made in order to remove ambiguities inherent in the usual symbols. These ambiguities arose largely as a result of a revised interpretation, in the new quantum theory, of certain coefficients used in the old quantum theory. Thus according to the old quantum theory, B_e was assumed the same as B_0, while according to the new theory, the two quantities are distinct (cf. Eqs. 3, 3a); the situation in regard to D_e, r_e and ω_e is similar.

Table I. *Symbols and notation used in Part I.*

(Prime)'; (Double Prime)''. These indices are used to distinguish the upper (') and the lower ('') of two energy levels involved in the production of an absorption or emission line. They may be attached to any symbol, coefficient, quantum number, or function which can differ for different states of a molecule (e.g. v', v'', B', B'', $U'(r)$, $U''(r)$, $x_e'\omega_e'$, $x_e''\omega_e''$, F'^v, F''^v, etc.).

$\mu = m_1m_2/(m_1+m_2)$. (m_1, m_2 are masses of nuclei, μ is "reduced mass").
I = moment of inertia $= \mu r^2$; $I_e = \mu r_e^2$.
r = distance between nuclei, in general; r_e = equilibrium value of r.
$\rho = r/r_e$; $\xi = (r/r_e) - 1$.
$U(r)$ = potential energy function for nuclear motions.
a^*, b^*, c^* = coefficients in $U(r)$, —cf. Eqs. (1).
D = energy of dissociation.
E = energy (ergs); E^r = rotational energy.
F = spectroscopic term $= E/hc$ (cm^{-1}).
F^{el}, F^v, F^r = electronic, vibrational, and rotational term (cf. Eq. 5).
$\Delta F^v(v_1, v_2) = F^v(v_2) - F^v(v_1)$ (cf. Eq. 18).
ν = spectroscopic frequency (in cm^{-1}) = difference of two terms $(F' - F'')$.
$\nu^{el} = F'^{el} - F''^{el}$; $\nu^v = F'^v - F''^v$; $\nu^r = F'^r - F''^r$ (cf. Eq. 7).
$\nu^0 = \nu^{el} + \nu^v$ (cf. Eq. 8); $\nu^{00} = \nu^{el} + \nu^v$ for $v' = 0$, $v'' = 0$ (cf. Eqs. 17, 17a).
$B_e = h/8\pi^2 c\mu r_e^2$.
$B \equiv B_v = B_e - \alpha(v+1/2) = B_0 - \alpha v$ (cf. Eq. 3a).
$\overline{B} = (B' + B'')/2$; $\overline{B}_0 = (B_0' + B_0'')/2$ (cf. Eqs. 15, 16a).
$C = B' - B''$; $C_0 = B_0' - B_0''$ (cf. Eqs. 15, 16a).
$D \equiv D_v = D_e + \beta(v+1/2) = D_0 + \beta v$ (cf. Eq. 4a); for D_e, cf. Eq. 4.
ω_e, $x_e\omega_e$, $y_e\omega_e$, α = coefficients in expression for F as a function of the quantum numbers (cf. Eqs. 5, 6).
$a = \omega_e(1-x)$; $b = x_e\omega_e$ (cf. Eq. 17).
$\tilde{a}$, $\tilde{b}$ = quantities differing slightly from a, b (cf. Eq. 17a).
$c\omega_e$ = frequency of mechanical vibration of nuclei for infinitesimal amplitudes about r_e.
v = vibrational quantum number.
K = rotational quantum number.
$K^* = [K(K+1)]^{1/2}$.

† Cf. important note on p. 115 in regard to further recommended changes.

* Also cf. the preliminary account in the Discussion on Molecular Spectra, Faraday Society Trans., pp. 628–633, 770–772, and Errata (Sept., 1929); also in book form (Gurney and Jackson, 1930).

P branch = negative branch ($K' = K'' - 1$).
R branch = positive branch ($K' = K'' + 1$).
M = ordinal number ($M = -K''$ in P branch; $M = K'' + 1$ in R branch).

Band; band-line; band-structure; band-system; band-head; electronic bands (cf. p. 68 and 72 for definitions).
Pure rotation bands; vibration-rotation bands (cf. pp. 83–4).
Band sequence; v' and v'' progressions (cf. p. 76).
Perturbations (cf. p. 74).

The following table is a supplement to the above Table I of Part I and is an index and summary of the symbols and notation introduced in Parts II*a* and II*b* but not used in Part I. It also serves to show the relation between the symbols used here—most of which are in accordance with an agreement recently arrived at by band spectroscopists,—and other symbols which are or have been in common use (*the latter are given in brackets*).

TABLE II. *Symbols and notation introduced in Parts IIa and IIb.*

A. Atoms.

n = principal, l = azimuthal, s = spin quantum number (individual electrons).
L = resultant orbital, S = resultant spin, J = resultant total angular momentum quantum number.
M_L, M_S, M = magnetic or electric quantum numbers associated with L, S, J, respectively; M_L, M_S, M = projections of L^*, S^*, J^*, respectively on axis of magnetic or electric field; $M_L h/2\pi$, etc., = corresponding angular momenta.
l^*, s^*, L^*, S^*, $J^* = [l(l+1)]^{1/2}$, $[s(s+1)]^{1/2}$, etc.; $l^* h/2\pi$, $s^* h/2\pi$, etc., are the angular momenta corresponding to l, s, etc.
s, p, d, f, · · · electrons: means $l = 0, 1, 2, 3, \cdots$
S, P, D, F, · · · states: means $L = 0, 1, 2, 3, \cdots$
A = coefficient of magnetic interaction of L^* and S^* (cf. Eqs. (22), (25), (26), and ref. 33).
co = Larmor precession frequency (magnetic field); g = Landé factor.
Weak and strong fields, definition (cf. ref. 30).
$c\omega_k$ = mechanical frequency associated with any quantum number k ($k = L$, J, M, K, etc.): cf. Eq. (23).

B. Molecules

A, ω_k, n, l, s, L, S, l^*, s^*, L^*, S^*, and s, p, d, f, · · · , mean the same as for atoms, except that l and L are usually not good quantum numbers, the corresponding angular momenta being in general not even nearly constant in magnitude.
J^a = same as *atomic* J; J^{a*} = resultant of L^* and S^* in cases where L^* and S^* are very strongly coupled.
Λ, Σ, Ω [usually called i_l or σ_k or σ_l, i_s or σ_s, and i or σ, respectively] = quantum numbers giving respectively the component of orbital, spin, and total electronic angular momentum parallel to the electric axis (cf. p. 93 for exact definitions); $\Lambda h/2\pi$, $\Sigma h/2\pi$, $\Omega h/2\pi$ = corresponding angular momenta.
Σ, Π, Δ, · · · states [formerly called S, P, D, · · · states] mean $\Lambda = 0, 1, 2, \cdots$; ${}^2\Sigma$, ${}^3\Pi_1$, ${}^4\Pi_{-\frac{1}{2}}$, etc., cf. p. 94.
L_{perp} (cf. p. 98), S_{perp} (cf. p. 105) = component of L^* or S^* perpendicular to electric axis, about which L^* or S^* is precessing; L_{perp} and S_{perp} are not quantum numbers.
$Gh/2\pi$ = total instantaneous component of electronic orbital angular momentum perpendicular to electric axis (cf. p. 98); G is not a quantum number.
$Nh/2\pi$ = instantaneous angular momentum of nuclei (cf. p. 99); N is not a quantum number.
$Oh/2\pi$ [formerly often called $mh/2\pi$] = total angular momentum perpendicular to electric axis, exclusive of spin in Hund's case b (cf. p. 97); O is not a quantum number.
$\rho h/2\pi$ [often called $\epsilon h/2\pi$] = mean value of projection of $Gh/2\pi$ on O axis = mean component of electronic orbital angular momentum along O.

R [Hund's p_r] = rotational quantum number in Hund's case d (cf. p. 101).
K [usually called j_k or p_l or m] = rotational quantum number, inclusive of electronic orbital angular momentum, but not of electronic spin (Hund's cases b, d).
J [formerly often called m] = quantum number of total angular momentum in all cases.
R^*, K^*, $J^* = [R(R+1)]^{1/2}$, etc.; $R^*h/2\pi$, etc., = angular momenta.
δ, ϵ, γ = small coefficients (cf. p. 100, and p. 107, Eq. (33)).
$B^*_{v,\Sigma}$ = (cf. Eq. (46)).
$\phi(K)$; $\phi_i(K)$ ($= \phi_a(K)$ or $\phi_b(K)$); $\phi_i(\Sigma, J)$; $\phi_i(K, J)$: small functions (cf. p. 100 et seq.; p. 105; p. 106; Eq. (32)).
$f(R, K\text{-}R)$; $f(K, J\text{-}K)$; $w(K, J\text{-}K)$: small functions (cf. Eq. (30); Eqs. (32-33)).
F_1, F_2, F_3, F_{1a}, F_{1b}, etc. = term designations in Hund's case b (cf. p. 112–13).
Hund's cases, definitions: Case a, p. 105; case b, p. 106; case b', pp. 97, 108; case c, p. 114; case d, p. 108; case d', pp. 101, 108; case e, p. 114; summary of Hund's cases, p. 115.
Core model (cf. p. 85).
Selection rules (cf. pp. 90, 96).
Vectors are denoted by bold-faced type.

INTRODUCTION

THE theory of the spectra of diatomic molecules has undergone very rapid development in the years since 1925. This has been especially true in respect to our knowledge (1) of the nature of electronic states and quantum numbers, and (2) of the connection between the electronic states of molecules and those of their component atoms on dissociation. Parts I–II, and Part III, of this review are respectively devoted primarily to the first and second of these topics, together with their application to the structure and interpretation of band spectra.

For the sake of simplicity and "Anschaulichkeit," the treatment in Parts I–II is in terms of the *old quantum theory* and repeatedly involves the use of models which, according to the new quantum theory, must not be taken too literally. So far as possible, however, the most essential new results of the new quantum theory,—especially energy relations,—are stated in the text, although their rigorous derivation is not given. In Part III, the new theory will be used more directly.

The possible energy values of a diatomic molecule can be expressed as functions of certain quantum numbers associated with motions of the electrons and nuclei. These quantum numbers usually fall naturally into three groups (*a*) a group which defines to a first approximation the energy, depending on the electronic motions, which the molecule would have if the nuclei could be held stationary (*b*) a single quantum number v which defines the state of vibration of the nuclei (*c*) a group of quantum numbers associated with the rotation of the nuclei and with the finer details of the electronic motions. In the simplest cases,—the so-called $^1\Sigma$ states,—the group (*c*) reduces to a single quantum number K belonging to the rotation of the nuclei[1,2], but in general there are interactions between electron motions

[1] K is usually called m or j. The reasons for using the symbol K will appear later (cf. pp. 97–106).

[2] Even in $^1\Sigma$ states, K does not represent *exclusively* nuclear rotation, because there is always a rapidly varying electronic orbital angular momentum (cf. J. H. Van Vleck and A.

and nuclear rotation which must be taken into account in the scheme of quantum numbers. It is toward an understanding of these interactions and their consequences for the structure of band spectra that Parts I–II are directed. But we shall begin (Part I) by reviewing the characteristic features, and their important variations, for the structure of a band spectrum corresponding to a transition from one $^1\Sigma$ electronic state[3] to another. In this way we shall be better able to appreciate the characteristic *new* structural features (Part II) which appear in bands involving *other* kinds of electron states.

PART I

Theory of Energy States of Molecule Regarded as Rotating Anharmonic Oscillator

Potential energy function. In molecular states of the $^1\Sigma$ type, the (average) angular momentum of the electron system is zero.[2] The motion of the nuclei is in this case essentially a pure rotation around their center of gravity, combined with a vibration along the line joining them. For each electronic state there is a function $U(r)$ which acts like a potential energy for the nuclear motions,[4] and which has a single minimum corresponding to an equilibrium value, r_e, of the distance r between the nuclei. If $U(r_e)$ is taken as zero, $U(r)$ gives the total energy of a molecule whose nuclei are momentarily at rest at distance r apart, and $U(\infty)$ is equal to the energy of dissociation D(cf. Fig.1). The following expansion for $U(r)$, expressed in terms of the quantities $\rho = r/r_e$ and $\xi = \rho - 1 = (r - r_e)/r_e$, has $U(r_e) = 0$ and is convergent for values of r not too far from r_e.[5]

$$U(\rho) = -a^*[-(1/2)+(1/\rho)-(1/2)\rho^2+b^*\xi^3+c^*\xi^4+\cdots] \qquad (1)$$

Eq. (1) may also be written as a series in ξ alone (Eq. 1a). Eq. 1a is readily obtained from Eq. (1) by expanding $1/\rho$ and $1/\rho^2$ in powers of ξ.

$$U(\xi) = a^*[\xi^2/2-(b^*+1)\xi^3-(c^*-3/2)\xi^4+\cdots] \qquad (1a)$$

Frank, Proc. Nat. Acad. Sci. **15**, 539, 1929). Corresponding to this fact, a term $B_e\overline{G^2}$ ought to be added to the rotational energy function in Eqs. 2a, 2b, or a term $B_v\overline{G^2}$ in Eq. 5 (cf. Part IIb, Eq. 29, with $\Lambda = 0$). $\overline{G^2}$ represents the mean value of the square of the component of electronic orbital angular momentum perpendicular to the line joining the nuclei. This differs from zero even when the average value of the electronic angular momentum vector is zero.

[3] A molecule is said to be in the same electronic state so long as the group of quantum numbers (a) remains constant, no matter how v and K may be varied. But if one or more of the quantum numbers of group (a) are changed, it is said to be in a different electronic state.

[4] Cf. E. C. Kemble, National Research Council Bulletin on Molecular Spectra in Gases, p. 293, for a discussion of the physical meaning of the function $U(r)$. (Kemble calls it $V(r)$).

[5] A. Kratzer, Zeits. f. Physik **3**, 289 (1920), but with designations of coefficients as given in appendix 15 of A. Sommerfeld's Atombau and Spektrallinien except that a^*, b^* and c^* are used here in place of a, b, and c of Sommerfeld. [In Sommerfeld's chapter, $-5b^2/2$ appears erroneously in the expression for $x_e\omega_e$ (cf. Eq.. (6) below) where $+5b^2/2$ should be used.] For further details of the energy expansion Eq. (5), cf. E. C. Kemble, Jour. Opt. Soc. Am. **12**, 1 (1926). For a form of $U(r)$ which is capable in many cases of representing this function for *all values* of r, cf. P. M. Morse, Phys. Rev. **34**, 57 (1929).

U (ξ) of Eq. (1a) is in practise considerably less rapidly convergent than $U(\rho)$ of Eq. (1), especially for polar molecules. $F(r) = -dU/dr$ is readily obtained from either of the above expressions, for example:

$$-F(\xi) = (1/r_e)(dU/d\xi) = (a^*/r_e)[\xi - 3(b^*+1)\xi^2 - 4(c^* - 3/2)\xi^3 + \cdots]. \qquad (1b)$$

Quantization of rotation. Let us review briefly the application of the quantum theory to the determination of the possible energy states corresponding to

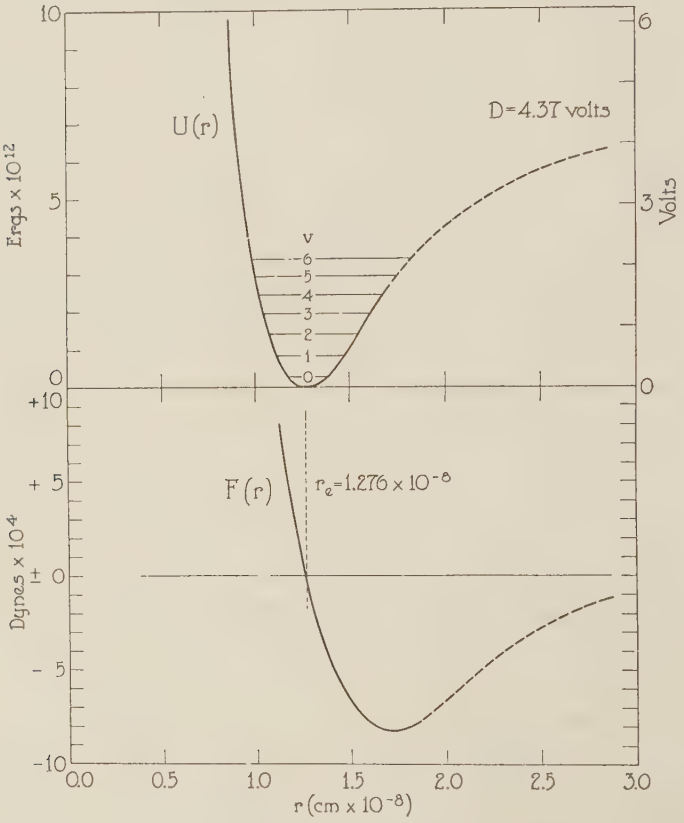


FIG. 1. Potential energy and force functions,—$U(r)$ and $F(r)$,—for nuclear motions, for the normal state of HCl; also a few vibrational energy levels. For r values below 1.8×10^{-8} cm, the curves are drawn in accordance with Eqs. 1, 1b; for larger r values, they have been approximately sketched in with the help of Morse's formula.[5]

the nuclear motions. First let us consider the (idealized) case where the motion is one of rotation without vibration. Let ϕ and P_ϕ respectively represent the angle through which the nuclei have rotated, and the angular momentum. Then, according to the old quantum theory, $\oint P_\phi d\phi = Kh$, where K is any non-negative integer, h is Planck's constant, and the integral is taken over a complete revolution. Since P_ϕ is a constant for a molecule in a definite state, the integral yields $2\pi P_\phi = Kh$, or $P_\phi = Kh/2\pi$. In the new quantum theory, this result is modified to $P_\phi = K^*h/2\pi$, where

$K^* = [K(K+1)]^{1/2}$. Our problem is now to express the energy of rotation in terms of K.

The energy of a rotating molecule is mainly kinetic energy, plus a small amount of potential energy resulting from a slight stretching by centrifugal force. If $\dot{\phi}$ is the angular velocity and I is the moment of inertia, the kinetic energy is $I\dot{\phi}^2/2$. If m_1 and m_2 are the masses of the two nuclei and r_1 and r_2 are their respective distances from the center of gravity, we have $I = m_1 r_1^2 + m_2 r_2^2$. The following treatment is simplified by rewriting I in terms of r and the "reduced mass" μ, defined[6] by $\mu = m_1 m_2/(m_1+m_2)$. We then have $I = \mu r^2$, as is easily shown by using the relations $r_1 + r_2 = r$ and (definition of center of gravity) $m_1 r_1 = m_2 r_2$. The potential energy is given to a sufficient approximation, for moderate speeds of rotation where $\xi \ll 1$, by $U = a^* \xi^2/2$ (cf. Eq. 1a). The corresponding force is given to a sufficient approximation by $-F = a^* \xi / r_e$ (cf. Eq. 1b). By setting $|F|$ equal to the centrifugal force $\mu r \dot{\phi}^2$, an expression for ξ in terms of $\dot{\phi}$ is obtained.[6a] Thus from $a^* \xi / r_e = \mu r \dot{\phi}^2$, putting r_e for r as a sufficient approximation and expressing $\dot{\phi}$ in terms of $P_\phi (P_\phi = I\dot{\phi} = \mu r^2 \dot{\phi} \sim \mu r_e^2 \dot{\phi})$, we get $\xi \sim P_\phi^2 / a^* \mu r_e^2$. For the total energy of rotation we now have

$$E^r = \mu r^2 \dot{\phi}^2/2 + a^* \xi^2/2 + \cdots = P_\phi^2/2\mu r^2 + P_\phi^4/2a^* \mu^2 r_e^4 + \cdots$$
$$= P_\phi^2/2\mu r_e^2 [1 - 2\xi + \cdots + P_\phi^2/a^* \mu r_e^2 + \cdots] = P_\phi^2/2\mu r_e^2 [1 - P_\phi^2/a^* \mu r_e^2 + \cdots]. \quad (2)$$

Substituting $P_\phi = K^* h/2\pi$,

$$E^r = K^{*2} h^2/8\pi^2 \mu r_e^2 - K^{*4} h^4/32\pi^4 a^* \mu^2 r_e^4 + \cdots. \quad (2a)$$

Adopting the abbreviations (c = velocity of light)

$$B_e = h/8\pi^2 \mu c r_e^2 \left[= 27.70 \times 10^{-40}/\mu r_e^2 \right] \quad (3)$$

$$D_e = -h^3/32\pi^4 a^* \mu^2 c r_e^4 \left[= -4B_e^3/\omega_e^2 (\text{cf. Eqs. 5, 6}) \right], \quad (4)$$

Eq. (2a) can be written in the form

$$F^r \equiv E^r/hc = B_e K^{*2} + D_e K^{*4} + \cdots = B_e K(K+1) + D_e K^2 (K+1)^2 + \cdots \quad (2b)$$

Simultaneous quantization of vibration and rotation. When the new quantum theory is applied to a molecule with zero average electronic angular momentum which is both vibrating and rotating,[7] the complete energy expression[2] is given by

[6] The correction for the masses of the electrons is too small to need to be included.

[6a] The centrifugal force is $m_1 r_1 \dot{\phi}^2$ for nucleus 1, $m_2 r_2 \dot{\phi}^2$ for nucleus 2. These are equal (since $m_1 r_1 = m_2 r_2$) and, as is easily shown, each is equal to $\mu r \dot{\phi}^2$.

[7] Old quantum theory, cf. Kratzer or Kemble, l.c., ref. 5; for relations such as Eq. (6) between coefficients in Eq. (5) and molecular constants, cf. Kratzer, l.c. New quantum theory, E. Fues, Ann. der Physik **80**, 367 (1926); **81**, 281 (1926), and A. Sommerfeld's Atombau und Spektrallinien, Wellenmechanischer Ergänzungsband, p. 24. It is probable that $(K+1/2)^2$ should be used rather than $K(K+1)$ in Eq. (5) and in similar equations, but the difference between the two expressions is negligible for practical purposes, and the form $K(K+1)$ gives simpler formulas for the frequencies of band lines.

$$F=F^{el}+F^{v}+F^{r}=F^{el}+(v+\tfrac{1}{2})\omega_e-(v+\tfrac{1}{2})^2x_e\omega_e+(v+\tfrac{1}{2})^3y_e\omega_e+\cdots$$
$$+B_vK(K+1)+D_vK^2(K+1)^2+\cdots \quad (5)$$

F in Eq. (5) means merely E/hc; throughout this review, for reasons of convenience in application to spectroscopic problems, we shall use *term*

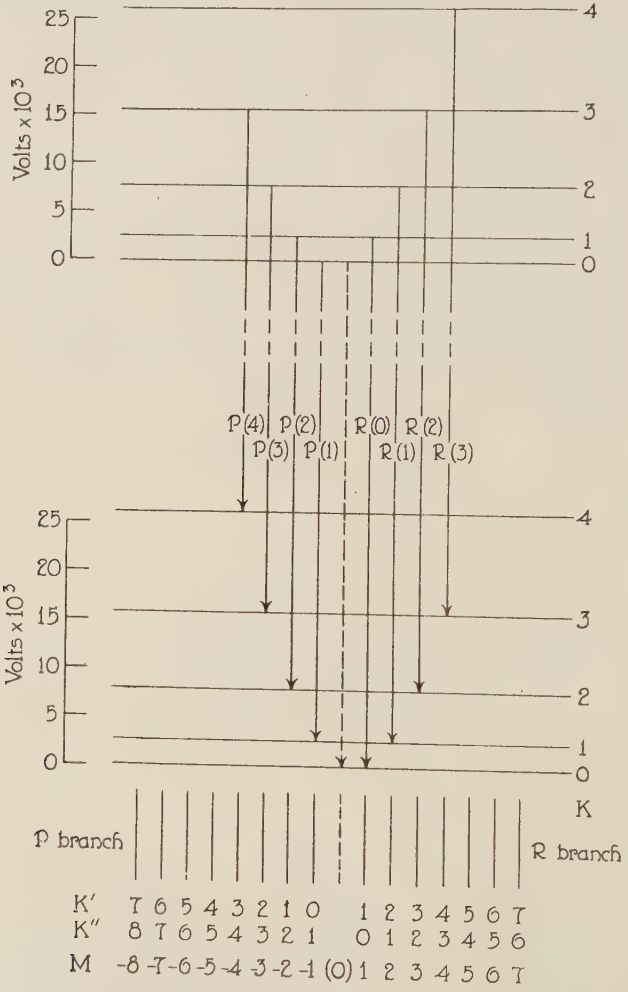


FIG. 2. Relation of rotational energy levels to band structure for the case $B'=B''$ (cf. Eq. 14); B'' is chosen equal to B_0 of normal HCl. The full vertical lines correspond to possible transitions between the two sets of energy levels. The figure shows the conventional method of designating band-lines, and its relation to the M numbering. The frequency which would correspond to $M=0$ (dotted lines in the figure) is not present in the spectrum; this is the so-called "missing line."

values (F) in place of corresponding energy values (E). In Eq. (5), F is, somewhat arbitrarily, represented as the sum of three terms, namely, a rotational term F^r which is a function of K (and of v and of the electronic state[3]), a vibrational term F^v which is a function of v (and of the electronic

state), and an electronic term which is a function of the electronic state only. The quantum numbers v and K can each take on positive integral values, beginning with 0. In Eq. (5), when $v=0$, $F^v=(1/2)\omega_e-(1/4)x_e\omega_e+\cdots\neq 0$; this represents the "zero-point vibrational energy," which cannot be removed from the molecule. It will be noted that F^r is the same in Eq. (5) as in Eq (2*b*), except for the substitution of B_v and D_v for B_e and D_e (cf. Eqs. 3*a* and 4*a* below).[2]

In Eq. (5), $\omega_e \gg x_e\omega_e \gg y_e\omega_e$ in practise, and $\omega_e \gg B_v \gg |D_v|$; for numerical examples, cf. pp. 67, 71, and 75. The "vibrational energy levels" have ordinarily a spacing which decreases slowly with v (cf. Fig. 1), while the "rotational energy levels" have a spacing which increases at first approximately as K^2 (cf. Fig. 2). As a result of $\omega_e \gg B_v$, the members of any set of rotational levels (fixed F^{el} and F^v, variable K) are very much more closely spaced than are the members of a set of vibrational levels (fixed F^{el}, $F^r=0$, variable v); this can be seen by a comparison of Figs. 1 and 2, with due regard to the scales used in these figures. The "electronic levels" (variable F^{el}, with F^v and $F^r=0$) are as a rule even more widely spaced than the vibrational levels, except in the case of highly excited electron orbits which are not very often met with in practise.

In Eq.(5), the coefficients are related theoretically as follows[7] to B_e and D_e of Eqs. (3) and (4) and to the molecular constants r_e, a^*, b^*, and c^* of Eqs. (1), (1*a*), and (1*b*).

$$B_v=B_e-\alpha(v+\tfrac{1}{2})+\cdots=B_0-\alpha v+\cdots, \text{ with } B_0=B_e-\alpha/2\ ;\ B_e\gg\alpha \quad (3a)$$

$$D_v=D_e+\beta(v+\tfrac{1}{2})+\cdots=D_0+\beta v\ ;\ B_e\gg|D_e|\gg\beta \quad (4a)$$

$$\omega_e=(1/2\pi c)(a^*/\mu r_e^2)^{1/2}\ ;\ \alpha=(6B_e^2/\omega_e)(2b^*+1)\ ;$$

$$x_e\omega_e=3B_e(1+5b^*+c^*+5b^{*2}/2). \quad (6)$$

In Eqs. (5) and (6), ω_e, times c, is the frequency of vibration of the nuclei for infinitesimal vibrations. Both α and $x_e\omega_e$ of Eqs. (3*a*), (5), and (6) are positive for all molecular states so far known; ordinarily or perhaps always in practise, $b^*>0$ and $c^*<0$.

Vibrational levels and $U(r)$ curve. When ω_e, $x_e\omega_e$, B_0, and α have been determined for any electronic state, as can be done by the analysis of band spectra connected with that state, the quantities r_e, a^*, b^*, and c^* can be computed, and the $F(r)$ and $U(r)$ curves can then be drawn, for r values not too far from r_e. If the dissociation energy D is known, the complete $U(r)$ curve can be approximately sketched in from $r<r_e$ to $r=\infty$. As we shall see later, such $U(r)$ curves are very useful in studying the behavior of the various electronic states of a molecule. Using the values $r_e=1.276\times 10^{-8}$, $a^*=82.87$, $b^*=0.173$ and $c^*=-0.316$, determined from the experimental coefficients[8] $\omega_e=2989.7$, $x_e\omega_e=51.90$, $B_e=10.58$ and $\alpha=0.303$, and the value $D=4.37$ electron-volts, the $U(r)$ and $F(r)$ curves for the normal state of HCl are sketched, using Eq. (1) and Eq. (1*b*), in Fig. 1. (It should be remarked that the true form of the $F(r)$ curve for HCl is very uncertain

[8] Cf. W. F. Colby, Phys. Rev. **34**, 53 (1929).

beyond about $r_e = 1.8 \times 10^{-8}$ cm, so that it is probably only qualitatively correct in Fig. 1.) A few of the vibrational energy levels of HCl are drawn on the $U(r)$ diagram in Fig. 1 in accordance with F^v of Eq. (5), assuming $K = 0$; the levels $v = 0$, 1, 2, and 3 have been observed, while the others are merely calculated.

P. M. Morse[5] has recently given a convenient formula by means of which a *complete* $U(r)$ curve can be computed if r_e, ω_e, and either $x_e\omega_e$ or D, are known. This is applicable only in cases where F^v of Eq. (5) can be fairly accurately represented, as is often but not always true, by $F^v = (v+1/2)\omega_e - (v+1/2)^2 x_e\omega_e$ without further terms. Morse's formula has been used in estimating the form of $U(r)$ for HCl in Fig. 1 for large r values.

Energy Levels and Spectrum

The wave-number ν of any line in the spectrum of a molecule can be obtained by taking the difference of two terms, one of higher energy (F') and one of lower energy (F''):

$$\nu = F' - F'' = (F'^{el} - F''^{el}) + (F'^{v} - F''^{v}) + (F'^{r} - F''^{r}) = \nu^{el} + \nu^{v} + \nu^{r}. \qquad (7)$$

Usually, but not always, $\nu^{el} \gg \nu^{v} \gg \nu^{r}$, corresponding to the relative energy level spacings of F^{el}, F^v, and F^r (cf. p. 67). All the spectrum lines associated with a definite pair of electronic states, hence with a definite F'^{el}, F''^{el} and ν^{el}, are collectively called a "*band system*." This is divided into limited groups of lines called *bands*, each band being associated with a definite pair of vibrational states and quantum numbers (v' and v'') and so with a definite F'^{v}, F''^{v}, and ν^{v}. The *lines* whose arrangement constitutes the *structure* of a band correspond to a variety of values of ν^r resulting from various possible pairs of values K', K'' of the rotational quantum number. The various lines are distributed over a limited ν region on both sides of the position, given by $\nu^{el} + \nu^{v}$, where $\nu^r = 0$.

We proceed now to a consideration of the structure of band systems which correspond to transitions between $^1\Sigma$ electronic states. We shall divide the discussion into two stages, first assuming ν^v = fixed, ν^r = variable, so as to obtain an understanding of the structure of individual bands, then later assuming $\nu^r = 0$, ν^v = variable so as to determine the arrangement of bands in a band system.

The bands of a band system with $\nu^{el} \neq 0$ are "*electronic bands*." There are also two other types of bands, for both of which $\nu^{el} = 0$, and which lie in the infra-red. These are the "pure rotation bands" and the "vibration-rotation bands." They are of relatively minor interest in the present review, but are considered briefly at the end of Part I.

Structure of Band Spectra of the Simplest Type ($^1\Sigma \rightarrow {}^1\Sigma$ Transitions)

Structure of individual electronic bands. In a *single band* of the electronic [illegible], ν^{el} and ν^{v} are fixed, while ν^r varies. In order to determine the form

which ν^r takes as a function of the quantum numbers, we make use in Eq. (7) of the detailed expressions for F' and F'' given by Eq. (5):

$$\nu=\nu^0+F'^r-F''^r=\nu^0+B_{v'}K'(K'+1)+D_{v'}K'^2(K'+1)^2$$
$$-B_{v''}K''(K''+1)-D_{v''}K''^2(K''+1)^2+\cdots. \quad (8)$$

Here ν^0 stands for $\nu^{el}+\nu^v$ of Eq. (7). Now there is an important selection rule which greatly limits the possible transitions and ν values: $(K'-K'')=\pm 1$ only. This is exactly analogous to the selection rule $\Delta l=\pm 1$ for the azimuthal quantum number in line spectra. If in Eq. (8) we put $K'=K''+1$ we get a set of frequencies given by the equation

$$\nu=\nu^0+(B'+B'')(K''+1)+(B'-B''+D'-D'')(K''+1)^2$$
$$+2(D'+D'')(K''+1)^3+(D'-D'')(K''+1)^4. \quad (9)$$

(For simplicity, the subscripts v' and v'' of the B's and D's have been dropped here). The series of lines given by Eq. (9) constitutes the so-called R or *positive* branch; in Eq. (9), K'' can take on the integral values 0, 1, 2, 3, ... Similarly, if we substitute $K'=K''-1$ in Eq. (8), we get the equation of the P or *negative* branch:

$$\nu=\nu^0-(B'+B'')K''+(B'-B''+D'-D'')K''^2-2(D'+D'')K''^3+(D'-D'')K''^4. \quad (10)$$

In Eq. (10), K'' can take on the integral values 1, 2, 3, ... ($K''=0$ is not possible, because it would imply $K'=-1$). Eqs. (9) and (10) can both be represented by the following Eq. (11) if we substitute $M=K''+1$ in Eq. (9) and $M=-K''$ in Eq. (10); in Eq. (11), $M=+1, +2, +3, \cdots$. give the lines of the R branch, $M=-1, -2, -3, \cdots$. the lines of the P branch. Eq. (11) shows that the two branches form a continuous series, except that a single central line $\nu=\nu^0$, corresponding to $M=0$, is missing.

$$\nu=\nu^0+(B'+B'')M+(B'-B''+D'-D'')M^2+2(D'+D'')M^3+(D'-D'')M^4. \quad (11)$$

The appearance of a band as recorded, say, on a photographic plate is governed by two factors (1) the arrangement of the lines as given by Eq. (11) and (2) the relative intensities of the lines. The first factor is subject to many variations depending on the relative and absolute values of B', B'', D', and D''. The second factor depends on (*a*) the initial distribution of molecules among rotational states and (*b*) the transition probabilities. In absorption spectra, and in *thermally excited* emission spectra, the initial distribution is that corresponding to thermal equilibrium at some temperature T. In practise, the observed intensity distribution, which depends of course on experimental conditions, is normally approximately of the type to be expected for thermal equilibrium, even when there is no reason to expect such equilibrium.[9] The theoretical intensity expressions, assuming

[9] Cf. R. T. Birge, Report on Molecular Spectra in Gases, p. 221; R. S. Mulliken, Phys. Rev. **29**, 401 (1927).

thermal equilibrium for the rotational energy distribution in the initial electronic state, are as follows for $^1\Sigma \rightarrow {}^1\Sigma$ transitions:[10]

Absorption:

$$\text{Int.} = C(K'+K''+1)e^{-E''^r/kT} = 2C \mid M \mid e^{-[B''M(M-1)+ \ldots]/kT}. \qquad (12)$$

Emission:

$$\text{Int.} = C(K'+K''+1)e^{-E''^r/kT} = 2C \mid M \mid e^{-[B'M(M+1)+ \ldots]/kT}. \qquad (13)$$

Eqs. (12) and (13) take into account both initial distribution and transition probabilities. Eqs. (12) and (13) show that, except for the exponential factor, the Mth line of the P branch should have the same intensity as the Mth line of the R branch. The exponential factor, however (because it involves M, not $|M|$) is for a given $|M|$ somewhat larger in the R than in the P branch for absorption spectra (Eq. 12) but is somewhat larger in the P than in the R branch for emission spectra (Eq. 13). All the features just discussed have been experimentally verified.[9]

We are now ready to consider how the structure of a $^1\Sigma \rightarrow {}^1\Sigma$ band may look in various actual cases. Let us first consider the simple case $B' = B''$, with D' and D'' negligibly small. Eq. (11) here simplifies to

$$\nu = \nu^0 + 2BM. \qquad (14)$$

The connection between the spacings of rotational levels (F^r of Eq. 5) and the spacings of lines in a band, for the case $B' = B''$ (Eq. 14), is shown in Fig. 2. In this figure the scale is so chosen that the spacings of the lower set of rotational levels agree with those experimentally determined for the $v'' = 0$ level of the normal electronic state of HCl. (The upper set of levels may then be considered to correspond to some imaginary excited state of HCl). Each band line in Fig. 2 may be regarded as the projection of one of the long vertical lines which connect pairs of energy levels. These vertical lines are spaced from left to right just in accordance with their length, corresponding to the fact that each spectrum frequency is proportional to the interval between the two energy levels involved in its production. The regular progression in the vertical lengths, which is the basis for the regular spacing of spectrum lines given by Eq. (14), should be apparent from the figure.

Fig. 2 also gives a comparison between the M numbering used here and the *conventional notation* which is commonly used for the designation of band lines. For our present descriptive purposes, the M numbering is more convenient than the other numbering, but the latter will be needed later, for more complicated types of bands and in connection with the analysis

[10] Cf. E. C. Kemble, Phys. Rev. **25**, 1 (1925); D. G. Bourgin, Phys. Rev. **29**, 794 (1927); and Eqs. (2) and (4) of Mulliken, Phys. Rev. **29**, 391 (1927). In Eqs. (12) and (13) above, C varies slightly with K, for large K values (cf. Kemble, l.c.).

To get from the first to the second form of the exponential in Eq. (12), one proceeds as follows. First one substitutes $E''^r = B''K''(K''+1)+ \cdots$, then, for the P branch, $K'' = -M$, for the R branch, $K''+1 = +M$. This gives for both branches the same result, $E''^r = B''M(M-1)+ \cdots$. A similar method applies in Eq. (13).

of band structure. In the conventional numbering, the value of K'' is given in parentheses, e.g. P (2) corresponds to $K'=1$, $K''=2$, while $R(4)$ corresponds to $K'=5$, $K''=4$.

Fig. 3 shows by means of illustrative examples how the band structure varies with B and T in emission spectra (in absorption spectra, the intensity relations of the P and R branches would be approximately reversed). For large B (small moment of inertia) the band has a very open structure, with few lines widely spaced and with the intensity maxima at small $|M|$ values but relatively far from ν^0. For small B everything is reversed. For small T, the number of strong lines, and the distance of the intensity maxima from ν^0, are relatively small. For large T there are relatively many lines,

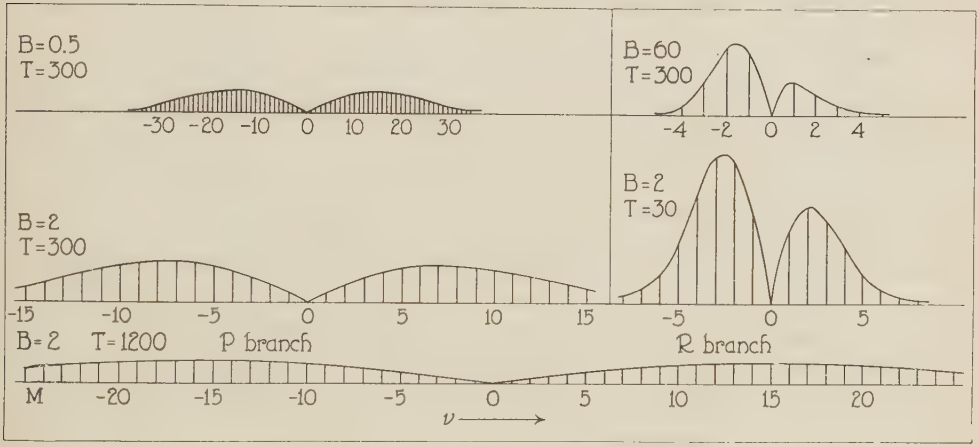


FIG. 3. Effect of numerical values of B and T on structure and intensity distribution (cf. Eqs. 13, 14), for emission bands having $B'=B''$ (headless bands). The numbers are M values. The heights of the lines represent their intensities. For the missing line ($M=0$), the intensity is zero. The law of intensity distribution has been emphasized by drawing envelopes of the individual intensities. All the diagrams are on the same scale, except for the case $B=60$, where both horizontal and vertical dimensions have been reduced to one-fourth what they should be.

and the intensity maxima are relatively far from ν^0. In practise, the largest B value is that of the normal state of $H_2(B_0=59.35)$.[10a] The various known states of He_2 all have $B_0\sim 7$; in the hydrides, B_0 varies from 21 in the normal state of HF to values such as 5.38 for the normal state of HgH and even lower in some other cases. Molecules like N_2, CO, NO have $B_0\sim 2$. The smallest B_0 value yet definitely determined is that of the excited $^1\Sigma$ level of the I_2 absorption bands ($B_0=0.029$). From these data and an examination of Fig. 3 it is evident that the resolving power needed to analyze the structure of a band varies enormously from one molecule to another.

The case $B'=B''$ just discussed is one which is only occasionally approximated in practise, except in infra-red bands with $\nu^{el}=0$ (cf. below). In

[10a] Cf. Birge, Hyman, and Jeppesen, Nature, 1930 (private communication from Professor Birge).

typical band spectra, $B'>B''$ or, somewhat more often, $B'<B''$. In either case Eq. (11) now becomes, still neglecting the D's,

$$\nu=\nu^0+2\bar{B}M+CM^2, \tag{15}$$

where $\bar{B}=(B'+B'')/2$ and $C=B'-B''$. Eq. (15) is plotted in Fig. 4 for several different pairs of values B', B'' (all near 2.0), for the case of emission

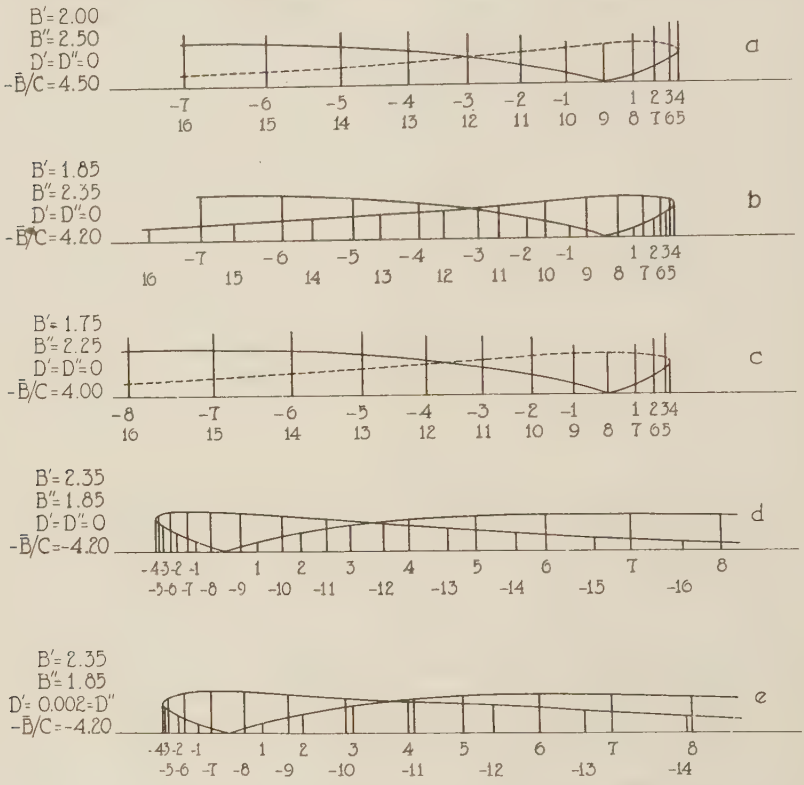


FIG. 4. Band structure types for $B'\neq B''$ (bands with heads) for emission bands with B values near 2 and $T=300°$ absolute (cf. Eqs. 13, 15). The direction of increasing frequency is toward the right. The numbers given under the lines are M values. Negative M values correspond to lines of the negative (P) branch, positive M values to lines of the positive (R) branch. The heights of the lines represent their intensities. In Figs. 4*a* and 4*c*, where lines of two series fall together, each height represents the sum of two intensities; but the separate intensities are also indicated by means of curves which represent the envelopes of the correct intensities for the separate series. [An error has inadvertently been made in the designation at the left of Fig. 4*e*, which should be corrected to read $D'=-0.002=D''$.]

spectra with an initial distribution corresponding to equilibrium at $T=300°$. The way in which the diagrams of Fig. 4 would need to be modified for a different T or a different order of magnitude of the B's can be judged from Fig. 3.

When $B'<B''$ (Fig. 4 *a*, *b*, *c*) the band has a "*head*" on the high frequency side, and is said to be degraded, or shaded, toward longer wave-lengths (or, loosely, toward the red); when $B'>B''$ (Fig. 4 *d*, *e*) there is a head on

the low frequency side, and the band is degraded toward shorter wavelengths (or, loosely, toward the violet). The head is usually the most conspicuous part of the band under low dispersion, since there is an accumulation of intensity there. The way in which the formation of a head is related to the relative spacings of the upper and lower rotational energy levels, as determined by B' and B'', can be seen by considering how the lengths of the vertical lines in Fig. 2 would be modified if $B'>B''$, or $B'<B''$, were assumed.

The approximate position of the head is given by taking $d\nu/dM$ in Eq. (15), and then setting $d\nu/dM=0$. This gives $M_{\text{head}}\sim-\overline{B}/C$; also, substituting in Eq. (15), we get ν_{head}:

$$M_{\text{head}}\sim-\overline{B}/C;\ \nu_{\text{head}}\sim\nu^0-\overline{B}^2/C \tag{16}$$

$\overline{B}/C$ may happen to be an integer (Fig. 4*c*) and in this case Eq. (16) is exact for M_{head}, but ordinarily, of course, $\overline{B}/C$ is not integral and the actual M_{head} is the whole number nearest to $-\overline{B}/C$. When $\overline{B}/C$ is a half-integer (Fig. 4*a*) two lines fall together at the head. Whenever $\overline{B}/C$ is either an integer or a half integer, the lines of the series returning from the head coincide exactly with the series of lines which is composed of the lines of the headless branch plus the lines of the head-forming branch before it reaches the head (cf. Figs. 4*a*, *c*). But in the ordinary typical case the lines of the returning series are regularly spaced *between* the remaining lines (cf. Figs. 4*b*, 4*d*). The relations mentioned in the last two sentences are true only so long as $2(D'+D'')M^3+(D'-D'')M^4$ of Eq. (11) is negligible. This is always the case for small M values, but when M becomes sufficiently large, the two series eventually cross, and as M increases further they cross repeatedly, as a result of the terms involving the D's and higher coefficients of Eq. (5). A comparison between Figs. 4*d* and 4*e* illustrates the effect of appreciable D's in bringing about the crossing of series; but it should be mentioned that the D's used in Fig. 4*e* are many times larger than would ever actually occur with the given B values.

In the last several paragraphs, we have seen how the structure of a band depends on the values of B' and B'', D' and D''. Now in a band-system, the structure is, slightly at least, different for every band, since B' and B'' are respectively functions of v' and v'' (cf. Eqs. 3*a*, 8, 11). Using Eq. (3*a*) we have

$B'=B_0'-\alpha'v'$ and $B''=B_0''-\alpha''v''$, where $\alpha'<<B_0'$ and $\alpha''<<B_0''$.

Hence in Eq. (16) we have

$2\overline{B}=B'+B''=B_0'+B_0''-\alpha'v'-\alpha''v''$, and $C=B'-B''=B_0'-B_0''-\alpha'v'+\alpha''v''$,

or

$$M_{\text{head}}\sim-\overline{B}/C=-(\overline{B}_0-\tfrac{1}{2}\alpha'v'-\tfrac{1}{2}\alpha''v'')/(C_0-\alpha'v'+\alpha''v'')\ ;$$
$$\nu_{\text{head}}\sim\nu^0-(\overline{B}_0-\tfrac{1}{2}\alpha'v'-\tfrac{1}{2}\alpha''v'')^2/(C_0-\alpha'v'+\alpha''v''). \tag{16a}$$

Thus $-\overline{B}/C$ varies from one band to another, so that in different bands of the same system any or all of the three types illustrated by Figs. 4*a*, *b*, *c* may occur.

When C_0 is small ($C_0 = B_0' - B_0''$), the variation of the B's with the v's is frequently enough to cause C to occur with *opposite sign* for different bands. Bands degraded toward long wave-lengths, headless bands, and bands degraded toward short wave-lengths may then all occur in a single band-system (example, violet CN bands).[11]

In addition to the variations already mentioned, there are sometimes irregularities of various kinds. Frequently there are "*perturbations*" in some of the lines of a band, i.e. displacements from expected position (and intensity). These are explained by the supposition that sometimes in a series of rotational levels, certain levels fail to obey the regular relation expressed by Eq. (5). Another type of irregularity, associated with certain forms of instability of the molecule, is a sudden breaking off of a series of lines, or a progressive broadening of successive lines and gradual fading out, beyond some value of M.

Another kind of departure from the relations discussed above can be expressed by introducing a small linear correction term[12] ϵK in Eq. (5). This disturbs Eqs. (9) and (10) in such a way that the P and R branches have a slight bodily relative shift, so that they can no longer be accurately represented by the single formula of Eq. (11). In Fig. 2 this shift would be equivalent to the use of a slightly different ν^0 (cf. Eq. 14) for the two branches.

Vibrational structure of band systems. Going back to Eqs. (5), (7) and (8), we now assume F'^r, F''^r and ν^r equal to zero in order to study the arrangement of bands in a system, taking the position $\nu^0 = \nu^{el} + \nu^v$ in each band as representative of the band. Neglecting terms in $(v+1/2)^3$, we have from Eqs. (5) and (7),

$$\begin{aligned}\nu^0 = \nu^{el} + \nu^v = &\left[\nu^{el} + \tfrac{1}{2}\omega_e' - \tfrac{1}{2}\omega_e'' - \tfrac{1}{4}x_e'\omega_e' + \tfrac{1}{4}x_e''\omega_e''\right] \\ &+ v'\omega_e'(1-x_e') - v'^2 x_e'\omega_e' - v''\omega_e''(1-x_e'') + v''^2 x_e''\omega_e'' \\ &= \nu^{00} + a'v' - b'v'^2 - a''v'' + b''v''^2.\end{aligned} \tag{17}$$

Usually this equation,— if necessary with additional terms in v^3 and so on,— is adequate to represent accurately the ν^0 values of the bands of a system. Occasionally, however, there are perturbations of some of the vibrational energy levels, so that the ν^0 values for some of the bands show irregular deviations from Eq. (17).

Just as the scale of the structure of individual bands varies with B, which depends on μ and r_e^2, so the scale of the spacing of bands in a band system varies with ω_e, which depends on μr_e^2 and a^* (cf. Eq. 6). Small ω_e results from small a^*, or from large μ or r_e; and conversely. Isotopic molecules are alike in respect to r_e and a^*, but differ in μ and so in ω_e. The largest

[11] Cf. F. A. Jenkins, Phys. Rev. **31**, 539 (1928).

[12] Cf. R. T. Birge, Report on Molecular Spectra in Gases, p. 171, etc. (CuH Bands). For a theoretical explanation, cf. Part II below, paragraph following Eq. (29).

ω_e known is 4371, for the normal state of H_2; next in size is 4037 for normal HF. Molecules like N_2 and CO have for their stable states values in the neighborhood of 1500–2300. Normal K_2 has $\omega_e = 92$; normal I_2 has $\omega_e = 128$.

In the analysis of a band system, the position ν_{head} (cf. Eq. 16a) is ordinarily measured, rather than the theoretically more important ν^0 of Eq. (17). In fact, it is usually not possible to determine the ν^0 positions without detailed analysis of the structure of a number of the bands. We need therefore a ν_{head} equation. Unless C is unusually small, the distance

v' \ v''	0	1	2	3	4	5	6	7	Ave. $\Delta F'^v$
0	20646 964	19682 950	18732 929	17803					
	862	862	862	857					861
1	21508 964	20544 950	19594 934	18660 918	17742				
	855	854	857	853	851				854
2	22363 965	21398 947	20451 938	19513 920	18593 904	17689			
	845	848	844	850	846	845			846
3	23208 902	22246 951	21295 932	20363 924	19439 905	18534 890	17644		
	832	840	841	838		837	837		838
4	24040 954	23086 950	22136 935	21201		19371 890	18481 873	17608	
		825	833	831			829	828	829
5		23911 942	22969 937	22032 921	21111		19310 874	18436	
			821	826	825			820	823
6			23790 932	22858 922	21936 908	21028		19256	
				818	817	818			818
7				23676 923	22753 907	21846 894	20952		
					811	811	810		811
8					23564 907	22657 895	21762 883	20879	
						803	804	807	805
9						23460 894	22566 880	21686	
Ave. $\Delta F''^v$	962	948	934	921	906	893	878		

$$\Delta F^v(v'', v''+1) = 962.6 - 14.03\, v'' = (\bar{a}'' - \bar{b}'') - 2\bar{b}''v''; \ \bar{a}'' = 969.6, \ \bar{b}'' = 7.01$$
$$\Delta F^v(v', v'+1) = 860.9 - 7.36\, v' = (\bar{a}' - \bar{b}') - 2\bar{b}'v'; \ \bar{a}' = 864.6, \ \bar{b}' = 3.68$$
$$\nu_{head} = 20646.4 + [864.6\, v' - 3.68\, v'^2] - [969.6\, v'' - 7.01\, v''^2]$$

FIG. 5. Most important part of v' v'' matrix diagram for frequencies of heads of AlO system, showing how the ΔF^v's (cf. Eqs. 18) can be determined, and from them the $\bar{a}$'s, $\bar{b}$'s, and the ν_{head} equation (cf. Eq. 17a). (The complete observed spectrum contains heads with v' up to 15 and v'' up to 12: cf. W. Mörikofer, Dissertation Basel, 1925). The frequencies given are experimental values, and for the weaker bands (especially the bands of the sequence $v' - v'' = 4$) are probably not very accurate, and so give rather irregular ΔF^v's. The ΔF^v and ν_{head} equations given do not correspond to a critical analysis of the data; Fig. 5 is intended to be illustrative rather than complete or exact.

$\nu_{\text{head}} - \nu^0$ does not vary much from band to band (cf. Eq. 16a). Hence the interval $\nu_{\text{head}} - \nu^0$ can be conveniently represented by an expansion of Eq. 16a in powers of v' and v''. This contains a constant term $-B_0^2/C_0$ and small terms in v', v'^2, v'', v''^2, $v'v''$, etc.[12a] Approximately, then, the heads of a band system should be given by an equation of the form

$$\nu_{\text{head}} = \nu^{00}{}_{\text{head}} + \bar{a}'v' - \bar{b}'v'^2 - \bar{a}''v'' + \bar{b}''v''^2 + \cdots. \qquad (17a)$$

[12a] Cf. R. S. Mulliken, Phys. Rev. **25**, 131 (1925), footnote 29.

Here $\nu^{00}{}_{\text{head}}$ means ν_{head} for the band $v'=0$, $v''=0$, while the $\bar{a}$'s and $\bar{b}$'s are constants differing just slightly from the a's and b's.

The ν^0 or ν_{head} values for the bands of a band system form according to Eqs. (17) and (17a) a two-dimensional array, and can be arranged in the form of a matrix. Fig. 5 shows such a matrix for a typical band system (AlO bands), using ν_{head} data as determined by experiment. Each band is characterized by two indices (v', v''); when one speaks of for instance the (3, 1) band, the numbers refer to these indices. There is no strict selection rule which limits the relative values of v' and v'', but the *intense* bands are limited in a way which depends largely on the relative values of ω_e' and ω_e'' (cf. discussion below).

Fig. 6 shows a reproduction of a photograph of the AlO bands, with the heads numbered according to v' and v''. The photograph shows well the grouping of the bands in "*sequences*," i.e. series of bands of constant $(v'-v'')$. It also shows well the $v''=0$ "*v' progression*" and the $v'=0$ "*v'' progression*." (In a v' progression, v'' is fixed, v' variable; in a v'' progression, v' is fixed, v'' variable). Fig. 6 shows also the relations,—neglecting the error introduced by using ν_{head} instead of ν^0 data,—between the positions of the various bands and the energy levels of the molecule. A careful study of the energy level diagram in connection with Fig. 5 and the photograph will make these relations clear.

From the ν_{head} data in Fig. 5, it is possible to determine the constants $\nu^{00}{}_{\text{head}}$, $\bar{a}'$, $\bar{b}'$, $\bar{a}''$, $\bar{b}''$ of Eq. (17a), hence approximately the corresponding constants of Eq. (17). One begins by taking *differences* of adjacent ν_{head} values (cf. Fig. 5). Such differences, which can be read off directly from the spectrum, must according to the theory be equal, except for the slight error caused by the variability of $\nu_{\text{head}}-\nu^0$ from band to band, to differences of the term values of adjacent vibrational energy levels of the molecule. For instance, except for the slight error caused by using ν_{head} data, $\nu(0,0)-\nu(0,1) = F''^{v}(1)-F''^{v}(0) \equiv \Delta F''^{v}(0,1)$. If the bands are correctly arranged in the matrix diagram, $\nu(1,0)-\nu(1,1)$ *also* should be equal to $\Delta F''^{v}(0,1)$, and we may therefore expect $\nu(0,0)-\nu(0,1)=\nu(1,0)-\nu(1,1)$ within experimental error. The weighted mean of these and other similar intervals $\nu(v_i', 0)-\nu(v_i', 1)$ may be taken as the best experimental value of $\Delta F''^{v}(0,1)$. Similarly $\Delta F''^{v}(1,2)$ can be obtained from the differences $\nu(v_i', 1)-\nu(v_i', 2)$. The above relations are exemplified in Fig. 5. In an analogous manner, we determine the $\Delta F'^{v}$'s: e.g. $\Delta F'^{v}(0,1)=\nu(1, v_i'')-\nu(0, v_i'')$, where $v_i''=0, 1, 2, \cdots$. The various $\Delta F'^{v}$ and $\Delta F''^{v}$ values from the AlO heads are given in Fig. 5. From Eq. (5) with $F^r=0$ we find that the ΔF^v's as obtained from ν^0 data should be quantities of the form

$$\left.\begin{aligned}\Delta F'^{v}(v', v'+1) &= \omega_e'(1-2x') - 2v'x_e'\omega_e' = (a'-b') - 2b'v' \\ \Delta F''^{v}(v'', v''+1) &= \omega_e''(1-2x'') - 2v''x_e''\omega_e'' = (a''-b'') - 2b''v''\end{aligned}\right\}. \quad (18)$$

The corresponding intervals from ν_{head} data are related in the same manner to the $\bar{a}$'s and $\bar{b}$'s. Thus the averaged intervals of Fig. 5 yield $\bar{a}'$, $\bar{b}'$, $\bar{a}''$, $\bar{b}''$, and, finally, a complete equation of the form of Eq. (17a).

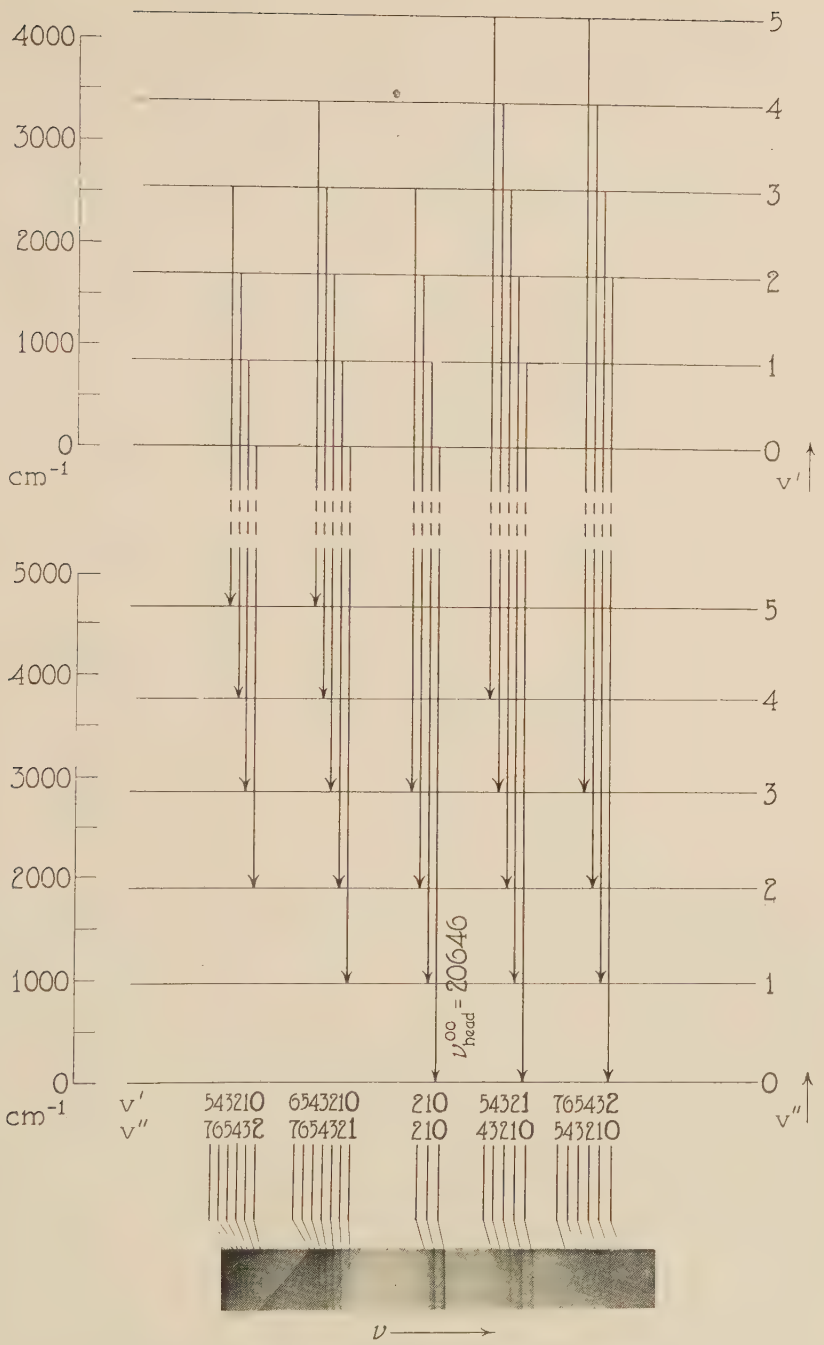


FIG. 6. Relation between vibrational energy levels and spectrum for the AlO band-system. At the bottom of the figure is a reproduction of a photograph showing the actual appearance of these bands as obtained with a prism spectrograph. Above this is a theoretical diagram on a uniform ν scale, with the v' and v'' values of each band. The observed spectrum consists mainly of five sequences of bands ($v'-v''=0, \pm 1, \pm 2$). Various v' and v'' progressions can be readily picked out from the reproduction or the theoretical ν diagram, e.g. the v'' progression with $v'=0$ and the v' progression with $v''=0$. The upper part of Fig. 6 shows a part of the vibrational energy levels of AlO, for each of the two electronic states involved in the AlO bands here considered. The spacings in each set of levels are given by a wave-number scale at the

The arrangement of bands in Fig. 5 serves to bring out certain criteria which must be fulfilled when a correct assignment of v' and v'' values has been made for a band system. These are: (1), if $b>0$ in Eq. (18), as is true in all cases yet known in practise,[12b] the experimental quantities corresponding to ΔF^v,—both $\Delta F'^v$ and $\Delta F''^v$,—must fulfill the inequalities ΔF^v $(0,1)>\Delta F^v$ $(1, 2)>\Delta F^v$ $(2, 3)\cdots$; (2), the matrix diagram (Fig. 5) should, in well developed band systems, be sharply bounded on two sides by two progressions

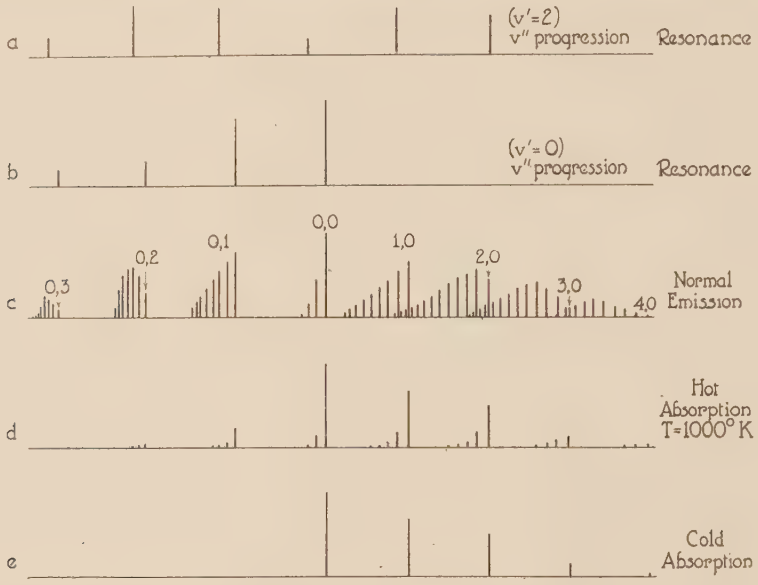


FIG. 7. Effect of initial distribution of molecules among vibrational states on vibrational intensity distribution, for a typical molecule. Each vertical line in the figure represents a band-head; its height is roughly proportional to the intensity of the band. In each band, the band-structure extends from the head toward longer wave-lengths, as in the reproduction in Fig. 6; the extent of this structure would vary, however, with the temperature (cf. Figs. 3, 4). All the diagrams are drawn in accordance with the known vibrational energy levels of the AlO molecule (cf. Fig. 5 for frequency data and Fig. 6 for energy levels and v' and v'' assignments). Absorption and resonance spectra have not yet been obtained experimentally for this molecule, so that the intensities in Figs. 7*a*, 7*b*, 7*d*, and 7*e* have been estimated largely from theoretical considerations; Fig. 7*c*, however, corresponds to the reproduced photograph in Fig. 6. It is of interest to compare the $v''=0$ and $v''=1$, v' progressions in the high-temperature absorption spectrum; the $v'=0$ band is strongest in the $v''=0$ progression, but the $v'=2$ band in the $v''=1$ progression. This difference, which is entirely one of transition probabilities, can be understood by a study of the Type III curves in Fig. 8.

of strong bands, namely the $v'=0$ progression and the $v''=0$ progression. In case the emitting or absorbing molecule has isotopes, there is also a third valuable criterion for the assignment of v values: isotopic molecules give band systems of identical structure but somewhat different scale, with the bands of all isotopes approximately coinciding near $\nu=\nu^{00}$.[13]

[12b] An exception appears to have been found in Li H (S. Nakamura, Zeits. f. Physik, **59**, 238, 1930.)

[13] Cf. R. S. Mulliken, Phys. Rev. **25**, 119 (1925), or F. W. Loomis, Report on Molecular Spectra in Gases, Chap. V.

We pass now to the question of the relative intensities of bands in a band system. In an emission spectrum, the bands which are developed depend very much on the conditions of excitation, since these govern the initial distribution of the molecules among vibrational levels. Fig. 7*c* represents diagrammatically the intensities of the AlO bands as obtained in emission by running an Al arc in air (cf. reproduction given in Fig. 6). A large range of v' values is present. So far as can be estimated by summation of the photographic intensities of bands originating from each v' value, there are in the arc more molecules with $v'=0$ than with $v'=1$, more with $v'=1$ than with $v'=2$, and so on. This is essentially a high-temperature thermal distribution, and is fairly typical for arc spectra.[14] With other types of excitation (electron impact, chemical reactions, excited atoms, etc.) other v' distributions often occur, e.g. the molecules may be confined to one or more, more or less sharply limited, low values of v'; or moderately high values of v' may be favored rather than low values; and so on.[14] In fluorescent spectra (emission spectra of molecules excited by light absorption) various v' distributions are possible, depending on the exciting wave-lengths and on the molecular constants. An interesting special case is that of "resonance spectra," that is, fluorescent spectra resulting from excitation by *monochromatic* light. Here we have a single value of v' (and of K'), and the emission spectrum consists of a series of bands (or rather, of particular lines of a series of bands) forming a v'' progression (cf. Fig. 7, *a* and *b*).

In absorption spectra, as in emission spectra, one has to consider the initial distribution among vibrational levels, but now this is a v'' distribution instead of a v' distribution. In a typical gas, such as CO, at room temperature, practically all the molecules have $v''=0$, and the absorption spectrum consists of a single v' progression with $v''=0$, as in Fig. 7*e*. This progression consists of bands all,—except the (0, 0) band itself,—on the high frequency side of the (0, 0) band. If the temperature is raised sufficiently, or (even at room temperature) if the molecule has a very small ω_e'' value (e.g., K_2 or I_2), the distribution is shifted so that an appreciable fraction of the molecules has $v''>0$. Additional v' progressions then appear (cf. Fig. 7*d*), so that the spectrum extends toward longer wave-lengths.

The relative intensities of the bands in a system are determined jointly by the initial distribution, as already considered, and by the relative *transition probabilities* as a function of $(v'-v'')$ and of v' or v''. Franck proposed a simple theory for approximately determining the transition probabilities; this was further developed by Condon, and has found many interesting applications in the hands of Franck, Birge and Sponer, and others.[15,16]

[14] Cf. R. S. Mulliken, Phys. Rev. **25**, 259; **26**, 1 (1925); R. T. Birge, Molecular Spectra in Gases, pp. 135–142; 221, 248–9; G. Herzberg, Ann. der Physik, **86**, 189 (1928); Zeits. f. Physik **49**, 512, 761 (1928); **52**, 815 (1929).

[15] For a review of this work, cf. H. Sponer, Ergebnisse der exakten Naturwissenschaften, Vol. VI, p. 75 (1927); also K. K. Darrow, Chem. Reviews, **5**, 451 (1928).

[16] Intensity distributions and Franck theory, cf. E. U. Condon, Phys. Rev. **28**, 1182 (1926).

Each electronic state of a molecule has, in general, a different $U(r)$ curve, and a different value of ω_e and of r_e. Suppose we consider the transfer of a molecule from one electronic state to a higher one by light absorption. The molecule is assumed initially in some definite vibrational state v''; the rotation of the molecule need not be considered. Franck argued that, to a first approximation, the light quantum acts directly only on the *electron* system

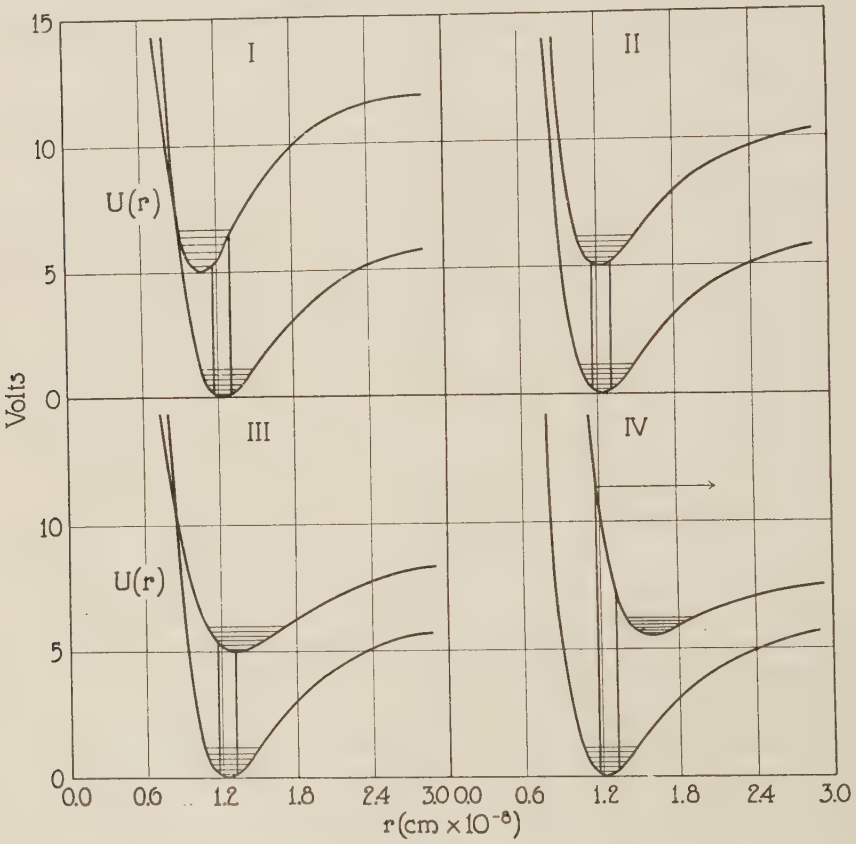


FIG. 8. $U(r)$ curves drawn according to Morse's formula,[5] and in accordance with the relation $r_e^3\omega_e = 3000 \times 10^{-24}$, for four typical cases. The curves are not those of any actual molecules, but the constants chosen are of approximately the same magnitude as is found for such molecules as O_2. Vertical lines drawn to the $U'(r)$ curves, from r_{min} and r_{max} of $v''=0$ of the various $U''(r)$ curves, serve to indicate the most probable v' values resulting from light absorption from $v''=0$. In Case IV, for r_{min} of $v''=0$, light absorption results in *dissociation*.

of the molecule, the inertia of the nuclei being too great for them to be directly affected. Thus the distance r between the nuclei, also the kinetic energy of the nuclei, should at the first instant be practically unchanged. This condition, in connection with the two $U(r)$ curves involved, suffices to determine the vibrational state after absorption.

Suppose we start with a molecule having $v''=0$. Its nuclear separation r may be thought of as oscillating between a maximum and a minimum

value (r_{max} and r_{min}). If it happens that $r \sim r_{max}$ or r_{min} when the light quantum is absorbed,—as is very probable, since the molecule spends most of its time near these values of r,—the correct state of vibration in the upper electronic state is obtained approximately by vertical projection from the $U''(r)$ curve to the $U'(r)$ curve. Thus in case II of Fig. 8, the molecule initially at r_{min} of $v''=0$ would find itself, after absorption of the quantum, at r_{min} of $v'=0$, and would go on with its vibration almost undisturbed by the electron jump. In case III, however, r_{min} of $v''=0$ is approximately the same as r_{min} of $v'=2$, and the molecule would probably vibrate with $v'=2$. In case IV, r_{min} of $v''=0$ corresponds to a point on the $U'(r)$ curve where $U'(r) > U'(\infty)$, and instead of a vibration, we should expect a flying apart of the nuclei (dissociation).

The procedure we have just been using is of course applicable for r_{max} as well as for r_{min}, for all values of v'', and for all types of $U(r)$ curves, in absorption spectra. It holds equally for emission spectra; in this case we project *downward* from the $U'(r)$ curve. In general, for any given initial v'' (or v') value there are *two* most probable values of v' (or v''), corresponding to the initial positions r_{min} and r_{max}. Hence for any given v'' (or v'), $(v'-v'')$ has two most probable values. Of course, other values of $v'-v''$ also occur, since the molecule spends part of its time between r_{min} and r_{max},[17] and also because the theory is only an approximate one. (That it is necessarily only approximate is evident for instance from the fact that vertical projection from r_{min} or r_{max} of a $U''(r)$ curve does not unless by chance hit the $U'(r)$ curve exactly at an energy corresponding to a permitted state of vibration.)

There are various empirical types of intensity distribution for the bands of a band system and, as Condon has shown,[16] these depend, in just the way given by Franck's theory, on the relative forms of the $U'(r)$ and $U''(r)$ curves. Each such type differs characteristically from the others not only in the intensity distribution with respect to v' and v'', but also in the *arrangement* of the bands, and in the appearance of the *band structure*. This threefold relationship can best be understood after an empirical relation which exists between the r_e and ω_e values has been stated. This is as follows: with only a few minor exceptions,[18] there is always, for the different electronic states of a single molecule, a *qualitatively inverse* relation between r_e and ω_e. Quantitatively, this has been expressed in the forms[19]

$$r_e^2\,\omega_e = \text{const.}; \text{ and, } r_e^3\,\omega_e \sim \text{Const.} \qquad (19)$$

[17] Suppose for example $r'=r_e''$ when the light quantum is absorbed. Here the molecule has kinetic energy equal to $E''^v - U''(r_e'')$ before the jump, and the projection must be made to a point above the $U'(r)$ curve by just this amount.

[18] In the case of two states with nearly equal r_e, the ω_e values are also nearly equal. Here it may happen that the larger ω_e goes with the larger r_e, contrary to the qualitative rule above stated. This is, however, obviously only a minor violation of the rule, and corresponds to the fact that no exact quantitative formulation of the rule can be given. It should also be emphasized that the rule is purely empirical; no theoretical reason for its validity is obvious.

[19] First form, cf. R. T. Birge, Phys. Rev. **25**, 240 (1925); R. Mecke, Zeits. f. Physik **32**, 823 (1925); second form, P. M. Morse, Phys. Rev. **34**, 57 (1929).

Both forms, especially the first, hold fairly well for *different states of a single molecule.* The second form, with *Const.* $= 3000 \times 10^{-24}$ cm^2, holds fairly well for a wide variety of states, even of *different molecules*, provided the masses of the two atoms are not too unequal (for unequal masses, *Const.* is larger; the largest values are for hydrides, where *Const.*$\sim$6000).

The statements of the previous paragraph can best be elucidated by considering the four typical pairs of $U(r)$ curves depicted in Fig. 8. In Type II of Fig. 8, where $r_e' = r_e''$, we expect $\omega_e' \sim \omega_e''$ according to Eq. (19). Because of $r_e' = r_e''$, we have $B_0' \sim B_0''$ (cf. Eqs. 3, 3*a*); but since $B' = B_0' - \alpha' v'$ and $B'' = B_0'' - \alpha'' v''$, we may have $B' \sim B''$, $B' > B''$ or $B' < B''$ in different bands, depending on the values of α', α'', v', and v''. Thus in the same band system we may find headless or practically headless bands, bands degraded toward long wave-lengths, and bands degraded toward short wave-lengths.—Because of $\omega_e' \sim \omega_e''$, the bands of each sequence are very crowded together or even almost superposed.[20]—Because $r_e' = r_e''$, the sequence $\Delta v = 0$ should according to the theory of Franck and Condon be very much stronger than any other sequence, with the sequences $\Delta v = \pm 1$ next in intensity, and so on. (A comparison with the typical band system shown in Fig. 6 will be helpful to the reader in considering the case here under discussion.)—The complete band system should then consist essentially of a strong central group of nearly superposed, headless or practically headless, bands comprising the sequence $(v', v'') = (0,0)$, (1, 1), (2, 2), etc., flanked by two much weaker sequences, likewise composed of nearly superposed bands, the shorter wave-length sequence comprising the bands (1, 0), (2, 1), (3, 2), ... , all degraded toward longer wave-lengths, and the longer wave-length sequence composed of bands (0, 1), (1, 2), (2, 3), · · · all degraded toward shorter wave-lengths.[21]

For $U(r)$ curves with $r_e' > r_e''$ by a good margin (cf. Type III of Fig. 8) we have $B' < B''$ for all bands of measurable intensity, so that these are all definitely degraded toward longer wave-lengths, like the AlO bands of Fig. 6. Accompanying $r_e' > r_e''$ we have $\omega_e' < \omega_e''$, and the members of each band sequence are now well spread out. At the same time, there are several strong sequences in the band system; the intensity distribution is of a type similar to that of the AlO bands in Fig. 6.

[20] If we had $a' = a''$, $b' = 0 = b''$ in Eq. (17), all the bands of a sequence would have their ν^0's precisely superposed. Actually, aside from the usual slight difference between a' and a'', the influence of b' and b'' causes successive members of a sequence to be separated. Usually b' and b'' differ greatly even if $a' \backsim a''$. An interesting case sometimes occurs in which (the series of ν^0's of) a sequence turns back on itself,—just like the lines of a band in the series which forms a head. This happens when, say, ω_e' is *slightly* greater than ω_e'' and x_e' is *decidedly* greater than x_e'', as is true in the violet CN bands (cf. F. A. Jenkins, l.c., ref. 11).

[21] If the sequences are *long*, as happens if ω_e is small or if the light source is such as to give large v' values, then the transition from bands degraded toward violet to headless bands and to bands degraded toward red may occur in a single sequence; this actually happens in the violet CN bands (cf. Jenkins, l.c., ref. 11). This is correlated (cf. Eq. 19) with the turning back of the sequences on themselves (ref. 20).

As the difference between ω_e' and ω_e'' increases (passage toward type IV of Fig. 8), each band sequence becomes more and more spread out and more and more overlaps neighboring sequences, until finally the classification into sequences loses its usefulness. At the same time, the intensity becomes distributed over a wider and wider range of v values. The arrangement of bands in the system becomes less obviously regular and the assignment of v' and v'' values more difficult.[22] With $U(r)$ curves as in type IV of Fig. 8, the band system includes bands of very large Δv, and is accompanied by a continuous spectrum corresponding in absorption bands to dissociation by light, or in emission bands to a union of atoms with light emission (examples: visible absorption and recombination spectra of the halogens). Finally, when r_e' and r_e'', and so ω_e' and ω_e'', are very different, we have *only* a continuous spectrum (examples, ultra-violet absorption spectra of the hydrogen and alkali halides).

When $r_e' < r_e''$ as in Type I of Fig. 8, we have a case very similar to that of Type III, except that now all the bands are degraded toward the violet. When the inequality of r_e' and r_e'' is larger, we have cases analogous to Type IV, except that here one has dissociation of a molecule accompanying light *emission* (example, continuous spectrum of H_2), and combination of atoms accompanying light absorption.

A more detailed treatment of the various types of intensity distribution among the bands of a system can be found elsewhere.[14,16]

Structure of $^1\Sigma \rightarrow {}^1\Sigma$ *bands: non-electronic bands.*[23] In Eq. (11), $\nu^0 = \nu^{el} + \nu^v$. If $\nu^{el} = 0$, we have infra-red bands of two types, (*a*) if $\nu^v \neq 0$, vibration-rotation bands; (*b*), if $\nu^v = 0$, pure rotation bands.

The vibration-rotation bands have the same type of structure as electronic bands (cf. Eqs. 11 and 15). The relative intensities of the band lines obey the same relations as electronic bands (cf. Eqs. 12 and 13.) But since $\nu^{el} = 0$, we have $B' = B_0 - \alpha v'$, $B'' = B_0 - \alpha v''$, with the *same* B_0 and α for both vibrational levels involved; hence, since α is small, we have $B' \sim B''$, approximating the cases shown in Fig. 3. But since B' here is always (slightly) less than B'', there is always a slight tendency to form a head on the R branch (cf. Eq. 16).

In vibration-rotation bands we have $x_e' \equiv x_e''$, $\omega_e' \equiv \omega_e''$, so that Eq. (17) becomes

$$\nu^0 = (v' - v'')\omega_e(1 - x_e) - (v'^2 - v''^2) x_e \omega_e . \qquad (20)$$

The relative intensities of different vibration-rotation bands are determined by entirely different factors (cf. below) than in the case of electronic bands. The strongest bands are those for which $v' = v'' + 1$ ("fundamental" bands). The "first harmonic" bands ($v' = v'' + 2$) are very much weaker, the second harmonic ($v' = v'' + 3$) very much weaker still, and so on. Since the infra-red bands

[22] Criterion (2) of page 78 now fails, since when the inequality of r_e' and r_e'', hence of ω_e' and ω_e'', is large, bands involving low values of both v' and v'', such as (0, 0), (0, 1), (0, 2), (1, 0), (2, 0), (1, 1), etc., are very weak or practically missing.

[23] Cf. W. F. Colby, Report on Molecular Spectra, Chap. III.

are ordinarily observed in absorption at room temperature, we usually have all the molecules in their normal electronic state and so have only a v' progression with $v''=0$. But at high temperatures something like a *band-system* would be possible, and in emission spectra we should expect to have one such system for each different electronic state.

The band intensities in vibration-rotation bands depend, according to the correspondence principle, on the variable part of the *electric moment* of the molecule. Homopolar molecules (H_2, N_2, O_2, etc.), on account of their zero dipole moment, have no vibration-rotation bands (electronic bands they *do* have).

In the pure rotation bands, $B'=B''$, $D'=D''$, and Eq. (11) reduces to

$$\nu=2BM+4DM^3 \tag{21}$$

M has only positive values here. Each band consists of a single series of nearly equally spaced lines. Such bands are best known for ($v''=0$ of the normal states of) the hydrogen halides. We shall not discuss here the relations which govern the intensities of the lines *within a band*, except to note that Eqs. (12) and (13) no longer apply. The intensity of any pure rotation band, *taken as a whole*, depends on the average dipole moment. Homopolar molecules, with zero dipole moment, have no pure rotation bands.

PART II

Introduction

In Part I we have considered, in summary, the important variations which can occur in the structure of a band spectrum associated with a transition between two $^1\Sigma$ electronic states. We shall now consider, in some detail, what new variations are possible when we have to deal with other than $^1\Sigma$ states. To begin with, it is important to note that these new variations do not directly touch ν^v, but only ν^{el} and ν^r, of Eq. (7). They consist mainly in new complications in the structure of individual bands, often accompanied by multiple values of ν^{el} analogous to atomic multiplets. It is convenient to approach this problem, following the method used by Hund in his fundamental paper on the subject,[24] by a study of a reasonable molecular model in terms of the old quantum theory. The immediate object, which will be followed in Part II, is to learn enough regarding energy relations and electronic quantum numbers to understand and interpret the various types of individual-band structure which occur in practise. Later, in Part III, we shall seek a more complete understanding of electronic quantum numbers and electronic states in molecules, and of their relation to the quantum numbers and states of their atomic dissociation products.

The discussion in Part II falls naturally into three stages: (*a*) a treatment of the problem of the electronic states of an imaginary limiting case which differs from real molecules in that the nuclei are assumed held stationary; (*b*) a consideration of the modifications required in part (*a*) in order

[24] F. Hund, Zeits. f. Physik **36**, 657 (1926).

to take into account the effect of the vibrations and rotations of the nuclei; (*c*) a study of the various types of bands resulting from transitions between electronic states of various types. Parts II*a* and II*b* are given below, while Part II*c* will appear at a later date.

II*a*. Theory of Molecular Electronic States for the Case of Fixed Nuclei

Comparison of molecule with atom in strong electric field. A molecule with nuclei assumed held fixed ("frozen molecule") can be regarded (1) as formed from two atoms or ions, each of which, when they come near, acts on the other with a large electric field; or (2) as obtained by dividing the nucleus of an atom into two parts and separating these somewhat, thus producing a modification of the original electric field of the single nucleus, equivalent to the superposition of a suitable correction field upon the original field. We shall here adopt the second point of view as the best means of reaching an understanding of the nature of the electronic states of molecules with nuclei held fixed. In an atom, the motion of any one electron can be treated as a motion in a spherically symmetrical but in general non-Coulombian electric field. The electrons, other than the one under consideration, are regarded as equivalent to a continuous *spherically symmetrical* cloud of negative electricity. (This assumption finds a large measure of justification in the new quantum theory.) This is the so-called core model (atom = core + chosen electron).

The corresponding model for the molecule with fixed nuclei consists in the nuclei plus an electron cloud having *axial* symmetry around a straight line passing through the nuclei. This molecular core-model can be imagined as built up from a suitably chosen atomic core-model by superposing on the central field of the latter a suitable, usually large, axially symmetrical correction field. This correction field must obviously change rather rapidly in both direction and magnitude, both along and at right angles to the axis. Such a correction field is of course quite different from the uniform unidirectional imposed field in the ordinary Stark effect for atoms. Nevertheless the existence of an axial symmetry in both cases is sufficient to make them qualitatively the same, so that the theory of the electronic states of a molecule with fixed nuclei is qualitatively identical with that of an atom in a strong electric field (Stark effect).

We therefore begin by a review of the results of the quantum theory for an atom in the absence of external fields, then, as a useful introduction to a study of the effect of electric fields, we consider briefly the effect of a uniform magnetic field on an atom (Zeeman effect). After that we pass to a more detailed consideration of atoms in axially symmetrical electric fields.

Quantum numbers and electron states of atoms.[25] In the absence of an external electric or magnetic field, the state of an atom can usually be de-

[25] For a detailed treatment of the theory of atomic spectra, reference may be made to F. Hund, Linienspektren und Periodisches System der Elemente (J. Springer, Berlin, 1927), or to the forthcoming book of L. Pauling and S. Goudsmit (McGraw Hill and Co., 1930).

scribed by giving quantum numbers n, l, and s for each extranuclear electron, and quantum numbers L, S, and J for the atom as a whole.[26] The principal quantum number n can take on values 1, 2, 3, $\cdots$, the azimuthal quantum number l can have values 0, 1, 2, 3, $\cdots$ (s, p, d, f, $\cdots$ electrons), while the electron spin quantum number s always has the value 1/2. In the old quantum theory, l for each electron is obtained by the quantization of the angular momentum P_l of the orbital motion of the electron in the field of the atom core (cf. above), all detailed interactions of the chosen electron with the core electrons being neglected, so that P_l can be treated as a constant. The quantum condition is $\int_0^{2\pi} P_l d\psi = lh$, which gives $P_l = lh/2\pi$. This last relation is replaced, according to the new quantum mechanics, by $P_l = l^* h/2\pi$, where $l^* = [l(l+1)]^{1/2}$. Similarly the angular momentum P_s of spin is given by $P_s = s^* h/2\pi = [s(s+1)]^{1/2} h/2\pi = \sqrt{3}h/4\pi$.

The quantum number L corresponds to an angular momentum $\boldsymbol{P}_L = \boldsymbol{L}^* h/2\pi$, where $L^* = [L(L+1)]^{1/2}$. [Bold-face type will be used throughout to indicate vectors.] $\boldsymbol{P}_L$ is the resultant of the individual $\boldsymbol{P}_l$'s. Similarly $\boldsymbol{P}_S = \boldsymbol{S}^* h/2\pi$, where $S^* = [S(S+1)]^{1/2}$, is the resultant of the $\boldsymbol{P}_s$'s. Most atoms contain a number of "*closed shells*." Each closed shell has $L=0$ and $S=0$ and therefore contributes nothing to the final resultant L and S of the atom. The latter are therefore determined solely by those electrons which are not in closed shells. If there is just one such electron, as in the alkali metals, then we have simply $L=l$ and $S=s$, where l and s refer to the valence electron.

If there are two electrons (1 and 2), not in closed shells, as in the excited states of the alkaline earth metals, the possible values of L are determined from l_1 and l_2 by the condition $L = l_1+l_2, l_1+l_2-1 \cdots |l_1-l_2|$. Corresponding to each such L value, a vector diagram can be drawn representing $\boldsymbol{L}^*$(or $\boldsymbol{P}_L$) as a resultant of $\boldsymbol{l}_1^*$ and $\boldsymbol{l}_2^*$ (or of $\boldsymbol{P}_{l_1}$ and $\boldsymbol{P}_{l_2}$). For S we have $S = s_1 \pm s_2 = 0$ or 1; these two possibilities occur for each of the possible L values. In each case $\boldsymbol{S}^*$ can be represented as a resultant of $\boldsymbol{s}_1^*$ and $\boldsymbol{s}_2^*$.—In case the two electrons are *equivalent* ($n_1 = n_2$ and $l_1 = l_2$), only part of the L, S combinations given by the rules just stated are possible; the allowed pairs of L and S values can always be determined by means of the Pauli exclusion principle.[25]

In an atom with three electrons (1, 2, 3) not in closed shells, the possible L and S values can best be obtained as follows. First we determine as before the possible L and S values for electrons 1 and 2, neglecting 3. In this way we get a variety of possible resultants L_{12} and two possible resultants $S_{12} = 0$ or 1. Now we add the third electron and get as possible values of the final L, $L = L_{12}+l_3$, $L_{12}+l_3-1$, $\cdots$, $|L_{12}-l_3|$, there being one such set of resultants for each L_{12}. In many cases this gives two or more equal L values which nevertheless correspond to physically different states. Since $s_3 = 1/2$, we have $S = 1/2$, from $S_{12} = 0$, and $S = 1/2$, 3/2 from $S_{12} = 1$. The two resultants $S = 1/2$ are physically different.—In case two, or all three, of the electrons are equivalent, the possible L, S combinations are fewer.

[26] The atomic notation used here follows the recommendations in the recent report by H. N. Russell, A. G. Shenstone, and L. A. Turner (Phys. Rev. **33**, 900, (1929)).

Atomic states are commonly classified according to their L values: S, P, D, F, $\cdots$ states respectively have $L=0, 1, 2, 3, \cdots$ The value of the spin S is also indicated, by means of a superscript number placed before the term symbol S, P, or D, etc. This number gives explicitly the "*multiplicity*" m and implicitly the value of $S(m=2S+1)$. Thus ^2S (to be read "doublet S", not "two S") means $L=0$, $S=1/2$, while ^1P("singlet P") means $L=1$, $S=0$; and so on.

For any given pair of values L, S, there are in general several values of the "inner" quantum number J, given by $J=L+S,\ L+S-1, \cdots |L-S|$. $J^*=[J(J+1)]^{1/2}$ is proportional to the resultant angular momentum P_J of the atom ($P_J=J^*h/2\pi$); each J^* may be considered as a resultant of the vectors L^* and S^*. When $L=0$, $J=S$; when $S=0$, $J=L$; in these cases it is not necessary to give J explicitly in the term symbol (examples, ^2S, ^1P); in other cases the value of J is given as a subscript, as for example in $^4P_{\frac{1}{2}}$ ($L=1$, $S=1\frac{1}{2}$, $J=\frac{1}{2}$) or in $^4P_{1\frac{1}{2}}$ ($L=1$, $S=1\frac{1}{2}$, $J=1\frac{1}{2}$).[27]

The energy of an atom usually depends in first approximation on the n and l values of the individual electrons, next on L and S, and finally on J. In an atomic multiplet level (fixed n's, l's, L, and S, variable J), different values of J correspond to different angles between the vectors L^* and S^*. Between L^* and S^* there is a magnetic interaction whose energy is proportional to the cosine of the angle between them. The torque exerted by L^* and S^* on each other causes them to precess (cf. Eq. (23)) around their resultant J^* which, since it represents the total angular momentum of the atom, maintains a fixed direction in space. Their mutual energy is given approximately[28] ($F=\text{term}=E/hc$) by

$$F-F_0=AL^*S^*\cos(L^*,S^*)=(\tfrac{1}{2})A\left[J(J+1)-L(L+1)-S(S+1)\right] \qquad (22)$$

F_0 is the "center of gravity" of the multiplet.[28a] Eq. (22) gives quantitative expression to the Landé interval rule, according to which, in a multiplet level, the spacing of any two sub-levels is proportional to the larger of the two J values concerned; e.g. in a ^4D level, the spacings of the intervals ($^4D_{3\frac{1}{2}}-{}^4D_{2\frac{1}{2}}$), ($^4D_{2\frac{1}{2}}-{}^4D_{1\frac{1}{2}}$), and ($^4D_{1\frac{1}{2}}-{}^4D_{\frac{1}{2}}$) are in the ratio $3\frac{1}{2}$: $2\frac{1}{2}$: $1\frac{1}{2}$.

It will be helpful to consider the application of Eq. (22) to ^2P and ^3P states. In ^2P states ($L=1$, $S=\frac{1}{2}$) we have $F-F_0$ equal to $(\frac{1}{2})A$ for the $^2P_{1\frac{1}{2}}$ level ($J=1\frac{1}{2}$), and equal to $-A$ for the $^2P_{\frac{1}{2}}$ level; this gives $\Delta F=1\frac{1}{2}\ A$ for the interval between the two levels. In ^3P (triplet P) states we have $F-F_0=A$

[27] The above quantum numbers and Eq. (22) are applicable to the so-called Russell-Saunders or normal type of coupling of the quantum vectors l and s. In some cases there is another type of coupling (reference 25) in which, instead of a resultant $L^*=\sum l^*$(vector) and a resultant $S^*=\sum s^*$ which then form a resultant J^*, we have a j and a j^*(resultant of l^* and s^*) for each electron (or at least for some electrons), and then $\sum j^*=J^*$.

[28] For further details, cf. Hund's Linienspektren (reference 25), Chaps. III-IV; S. Goudsmit, Phys. Rev. **31**, 946, (1928); S. Goudsmit and C. J. Humphreys, Phys. Rev. **31**, 960 (1928).

[28a] The terms "multiplet" and "multiplet level" are here used, interchangeably, to designate a group of *energy levels* differing only in J (atoms), or in Σ (molecules). The term "multiplet" is also (more appropriately) used to designate a group of *spectrum lines* derived from a pair of multiplet levels.

for $^3P_2$, $-A$ for $^3P_1$ and $-2A$ for $^3P_0$; this gives $\Delta F = 3A$ for the over-all width ($^3P_2 - {}^3P_0$), and also fulfills the Landé rule for the separate intervals.

Mechanical frequency equation. There is in the classical mechanics an important theorem regarding frequencies of motion. This theorem yields in the old quantum theory the following relation for the frequency $c\omega_k$ associated with any quantum number k (we use $c\omega_k$, rather than ω_k, for frequency, because throughout this review the symbol ω is used to represent a quantity measured in cm^{-1}).

$$c\omega_k = (1/h)(\,|\,\partial E/\partial k\,|\,) = c\,|\,\partial F/\partial k\,| \tag{23}$$

For example, the frequency of precession associated with J is that of $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ around their resultant $\boldsymbol{J}^*$; this is obtained by taking $c(\partial F/\partial J)$ in Eq. (22), which gives $cA(J+\frac{1}{2})$.

The reader should be cautioned not to attach too exact a meaning to Eq. (23), especially for low quantum numbers. A quantitative use of Eq. (23) is not permissible in the new quantum mechanics. Qualitatively, however, Eq. (23) is very useful in studying the behavior of vector models such as are used here.

Atoms in magnetic fields. When a uniform external magnetic field is imposed on an atom, $\boldsymbol{J}^*$, instead of remaining fixed in space, precesses around the direction of the field, making with the latter a constant angle θ. This angle is limited to certain values given by $\cos\theta = M/J^*$, where M is a "magnetic" quantum number, which can take on the values $J, J-1, \cdots, -J$. $\boldsymbol{M}h/2\pi$—not $\boldsymbol{M}^*h/2\pi$—represents the component, parallel to the field, of the angular momentum $\boldsymbol{J}^*h/2\pi$. (An analogous relation holds for all angular momenta which are obtained by *projection* on a relatively fixed axis.)

The precession of $\boldsymbol{J}^*$ around the field results from the fact that the latter exerts a torque on $\boldsymbol{L}^*$ and on $\boldsymbol{S}^*$. The effect of the field is, however, different for $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$. In a field of strength H, $\boldsymbol{L}^*$ tends to precess with the Larmor frequency $co = eh/4\pi m$, but $\boldsymbol{S}^*$ with a frequency $2co$ (e and m are the charge and mass of the electron). The actual precession of $\boldsymbol{J}^*$ occurs with a compromise frequency gco (cf. below, after Eq. (24)), where g is a factor depending on the values of L, S, and J. When $S=0$ and $L>0$ as in the singlet states ^1P, ^1D, etc., $g=1$; when $L=0$ and $S>0$ as in the S states ^2S, ^3S, etc., $g=2$; in a multiplet state with $L>0$, g is in general different for each value of J. The *energy* of the atom is now given[29] by

$$F - F_0 = (\tfrac{1}{2})A\,[J(J+1) - L(L+1) - S(S+1)] + Mgo. \tag{24}$$

Taking $c\,|\partial F/\partial M\,|$ in Eq. (24), in accordance with Eq. (23), we get for the frequency associated with $\boldsymbol{M}$ the value cgo; this represents the frequency around the field.

The compromise precession of $\boldsymbol{J}^*$ around the field exists only so long as the magnetic interaction between $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ is strong enough to preserve their resultant $\boldsymbol{J}^*$ from the disruptive influence resulting from the tendency

[29] Cf. Hund's Linienspektren (reference 25), Chaps. III-IV.

of the external field to make L^* and S^* precess at different rates. This is true so long as the frequency $cA(J+\frac{1}{2})$ of precession of L^* and S^* about J^* is large compared with the frequency gco of J^* around the external field (it should be remembered that o is proportional to the field strength). If the field is sufficiently strong, J^* is no longer constant and has no quantum significance,[29a] while L^* and S^* precess separately around the field, giving magnetic quantum numbers M_L and M_S ($M_L = L, L-1, \cdots -L$, $M_S = S, S-1, \cdots -S$). The corresponding angular momenta are $M_L h/2\pi$ and $M_S h/2\pi$ (not $M_L{}^* h/2\pi$ and $M_S{}^* h/2\pi$,—cf. above). The total energy is given by F_0 of Eqs. (22) and (24), plus the magnetic energy oM_L of L^* in the field, the

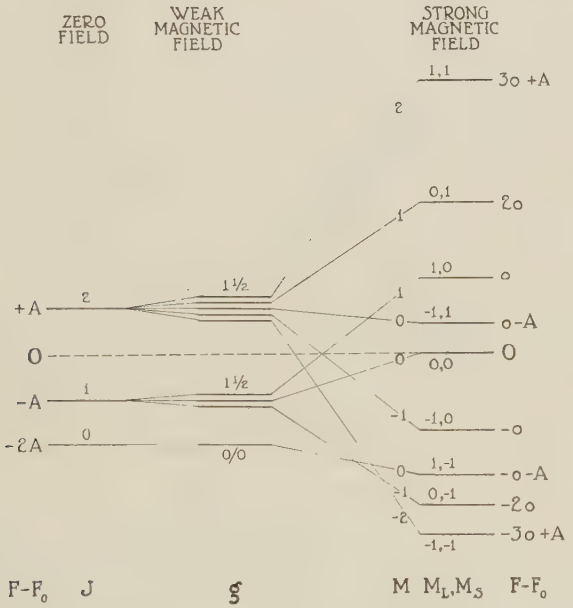


Fig. 9. Energy levels and quantum numbers for a ^3P atom in weak and strong magnetic fields. Energies are expressed by term values F measured from F_0; $F-F_0$ is expressed in terms of A, o, g, M, M_L, and M_S of Eqs. (24) and (25).

magnetic energy $2oM_S$ of S^* in the field, and the mutual energy AM_LM_S of S^* and L^* (the explanation of this mutual energy term will be clear from the discussion, preceding Eq. (26), of the analogous term in that equation; the "energies" mentioned are of course really all E/hc values):

$$F-F_0 = oM_L + 2oM_S + AM_LM_S. \tag{25}$$

It will be noted that the frequencies of L^* and S^* around the field are correctly given as co and $2co$ respectively, by the application of Eq. (23) to Eq. (25).

[29a] On account of the different rates of precession of L^* and S^* in this case, their resultant J^* varies rapidly in direction and magnitude, and J is no longer a quantum number.

In the passage from a "weak" to a "strong" field[30] (Paschen-Back effect), the number of magnetic levels remains unchanged; for each M value in a weak field, there is a corresponding M value (here $M = M_L + M_S$) in a strong field. Fig. 9 shows the transition from weak to strong field [30] for a normal ^3P state ($A > 0$ in Eqs. (22), (25)). For weak fields the spacings of the levels are governed by Eqs. (22) and (24), for strong fields by Eq. (25); the proper connections between weak and strong fields are qualitatively established by means of the two rules[29] (*a*) M remains constant during the transition; (*b*) in case there are several levels with the same M, the connections are drawn in such a way that no two levels with the same M cross in the passage from a weak to a strong field.

Selection rules for atoms. In transitions between different states of an atom, changes in some of the quantum numbers are limited by selection rules. In the absence of an external field these are: $\Delta l = \pm 1$ for one electron, $\Delta l = 0$ or ± 2 for a second electron (but usually the quantum numbers change for only one electron at a time); $\Delta L = 0, \pm 1$; $\Delta J = 0, \pm 1$.

In a weak magnetic field, the same selection rules hold, and in addition there is $\Delta M = 0, \pm 1$. In the passage to a strong field, the selection rule for J loses its strictness in proportion as J loses its significance. In a strong field, the selection rules for l and L remain strictly applicable, and in addition $\Delta M_L = 0, \pm 1, \Delta M_S = 0$ hold.

Atoms in electric fields. The effects of an axially symmetrical electric field[31] are similar to those of a uniform magnetic field, but differ considerably in some respects. Such a field,—whether or not it is uniform,—tends to make L^* precess[32] around it with a frequency whose magnitude for a given

[30] The terms "weak" and "strong" field are ordinarily used in a purely relative sense A field is strong or weak according as it is or is not able to break down the coupling of L^* and S^* which gives J^*. Of course the ideal weak or strong field is only *approximated* in practice. The energy levels under the respective headings "weak" and "strong" fields in Figs. 9 and 10 are drawn as if for *ideal* weak and strong fields, although because of space limitations the A values and field strengths assumed in the figures would actually correspond to very poor approximations to these limiting cases.

[30a] In the strong electric field case with $M_L = 0$, or with $A = 0$ in Eq. (27), there is no magnetic field to orient the spin S^*. The latter therefore acts completely independently of the rest of the molecule so that, if external influences are absent, the orientation of its axis (which maintains a fixed direction in space) is purely random. Thus there is no sense in defining M_S, and it is at any rate not a quantum number. (One *can* define M_S if one wishes, although in a real molecule this quantity varies very rapidly in direction and magnitude because of the fact that the molecular electric axis rotates, while the spin axis stands still). It should, however, be pointed out that in passing from the *ideal case* of a strong electric field, with $M_L = 0$, to the corresponding energy level in a weak electric field, the spin becomes coupled to the electric axis provided $L > 0$ (cf. text, p. 95 and ref. 35).

[31] Cf. Reference 25, Hund's Spektrallinien, p. 76–8 and references there given; also R. S. Mulliken, Phys. Rev. **33**, 746 (1929).

[32] Commonly in molecules the action of the field is even stronger, so that it breaks down the coupling of the l^*'s to give L^*. In this case, each l^* precesses separately around the field axis. This does not, however, alter the essential qualitative features which exist when L^* precesses around the electric axis. The theory of the latter case is therefore adequate, even if not entirely appropriate, for the present treatment of molecular electron states. In Part

field depends very much on the character of the electron orbits associated with L^*. An electric field exerts no direct influence whatever on S^*, since S^* corresponds to a so-to-speak disembodied magnetic moment. If the coupling between L^* and S^* is strong, however, ("weak" electric field), their resultant J^* orients itself, as in a weak magnetic field, with $\cos\theta = M/J^*$, where M takes any one of the values $J, J-1, \cdots, -J$. M measures the component of angular momentum parallel to the electric axis.

If the action of the external electric field on L^* is strong compared with the interaction of L^* and S^*, L^* precesses independently of S^* around the field axis,[32] giving $M_L = L, L-1, \cdots, -L$. Although S^* is not acted on by the electric field, it *is* acted on by the magnetic field of L^*, even though it is no longer coupled with L^* to give a constant resultant J^*. The magnetic field of L^* may be resolved into a fixed component directed along the electric axis, and a precessing component perpendicular to the axis but having an average value zero because of its precession. S^* precesses around the direction of the average magnetic field of L^*; but since the latter is parallel to the electric axis, S^* itself precesses, like L^* but for a different reason, about that axis; it gives thus a quantum number M_S whose values range from S to $-S$. The total angular momentum parallel to the electric axis is $Mh/2\pi$, where the quantum number M is given by $M = M_L + M_S$.

The energy of S^* in the field of L^* is equal, as in Eq. (22), and with the same value of A, to $AL^*S^*\cos(L^*, S^*)$, except that here one must take the *average* cosine, $\overline{\cos(L^*, S^*)}$; this is equal[33] to $M_L M_S / L^*S^*$. That is,[31]

$$F - F_0 = F(M_L) + AL^*S^* \overline{\cos(L^*, S^*)} = F(M_L) + AM_L M_S. \qquad (26)$$

In the special case $M_L = 0$, the energy of interaction of L^* and S^* is zero; here S^* has nothing to orient it and is entirely free, so that M_S has no quantum significance.[30a]

In Eq. (26), $F(M_L)$ corresponds to the energy of L^* in the electric field, expressed as a function of M_L. $F(M_L)$ has the important property that $F(M_L) = F(-M_L)$, i.e. the energy of orientation of L^* in an axially symmetrical electric field depends only on the *magnitude* of M_L. Why this is true can best be understood by thinking of the simple case of an atom with one valence electron, all other electrons being in closed shells. Here L^* is merely l^* of the valence electron, and M_L becomes m_l. Neglecting the spin, whose energy is taken care of by $AM_L M_S$ of Eq. (26), the forces acting on the electron in its orbital motion are purely electrostatic. Hence the energy asso-

---

III we shall consider the various coupling cases more thoroughly, including the case of fields so strong as to break down even the quantization of the individual l's, as in the Stark effect of the H atom (*linear* Stark effect). (Cases in which J^*, L^*, or the l^*'s precess all fall under the heading of *quadratic* Stark effects.)

[33] That $L^*S^*\overline{\cos(L^*, S^*)} = M_L M_S$ is most easily seen by writing $L^*S^*\overline{\cos(L^*, S^*)} = \overline{L^* \cdot S^*}$, that is, as the mean value of the scalar product of the vectors L^* and S^*. Resolving each of these vectors into components perpendicular and parallel to the electric axis, and noting that each perpendicular component averages zero, the stated result is readily obtained.

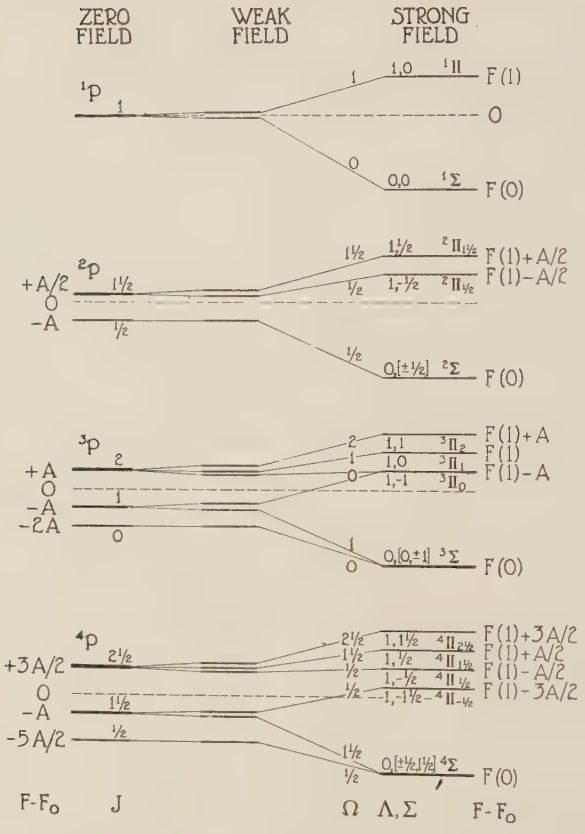


Fig. 10. Energy levels and quantum numbers for ^1P, ^2P, ^3P, and ^4P atoms in weak and strong electric fields.[30] Energies are expressed by giving $F-F_0$ values in accordance with Eq. (27). For zero field, we have $L=1$, with various S and J values. For weak fields, there is also Ω. In strong fields, there is Λ, which in the present examples has the two values 0 and 1 with corresponding $F(\Lambda)$ values,—cf. Eq. (27),—$F(0)$ and $F(1)$. In strong fields, if $\Lambda>0$, there is also a Σ; if $\Lambda=0$, Σ does not exist, but certain virtual values of Σ can be assigned (these are given in brackets in the figure); there is also $\Omega=|\Lambda+\Sigma|$. The quantities $F(0)$ and $F(1)$ have been calculated according to the equation $F(\Lambda)=a[3\Lambda^2-L(L+1)]$. The functions $F(\Omega)$ for weak fields have been calculated according to $F(\Omega)=b_J[3\Omega^2-J(J+1)]$, with a different value of b_J for each J. These equations (cf. W. Pauli, reference 34, p. 252) have no quantitative significance, but are of a form qualitatively suited to represent the effect produced on the terms of an atom by dividing its nucleus into two parts and separating these somewhat (cf. calculations of Morse and Stueckelberg, Phys. Rev. **33**, 932, 1929, in regard to the effect on the He^+ terms of splitting the He nucleus). In the equations given, a and b_J increase with the applied field, or with the degree of separation of the nucleus. Because $F(\Omega)$, like $F(\Lambda)$, depends really on the energy of L^* in the field, b_J, if a is positive, should have a positive value when L^* and J^* make an angle near 0 or π but a negative value when the angle between L^* and J^* is near $\pi/2$. In Fig. 10, a negative value of b_J has accordingly been assumed for $^3P_1$ and $^4P_{1\frac{1}{2}}$ but a positive value in other cases.

ciated with the valence electron, in an orbit of given n and l, depends only on the inclination of the plane of the orbit with reference to the electric axis, and not on the direction of rotation of the electron in its orbit. Now $\pm m_l$ correspond to the same inclination of the orbital plane although they differ in respect to the direction of rotation. Hence the energy is the same for $\pm m_l$. (In a magnetic field, where the magnetic energy depends on the vector velocity, the relations are entirely different: the change from $+m_l$ to $-m_l$ reverses the sign of the energy.) It is easy to see that the same results hold for M_L in the case where $\boldsymbol{L}^*$ is the resultant of several l*'s.

Further, we have $F(M)=F(-M)$, for both strong and weak fields. For strong fields, this is clear from Eq. (26), since $F-F_0$ is unchanged in the latter if we simultaneously change the signs of M_L and M_S, as is necessary to reverse the sign of M. For weak fields also, where $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ are coupled to give $\boldsymbol{J}^*$, a little consideration shows that the energy of orientation of $\boldsymbol{J}^*$ in the field, depending entirely on the $\boldsymbol{L}^*$ part of $\boldsymbol{J}^*$, is the same for $+M$ as for $-M$.

Fig. 10 shows qualitatively the energy levels for ^1P, ^2P, ^3P, and ^4P atomic states in axially symmetrical electric fields of various intensities.[30] The correlation between the levels in strong and weak fields is effected by means of the M rules stated above in connection with Fig. 9, applied with due regard for the fact that each level with $|M|>0$ is a level of double statistical weight, corresponding to $\pm M$. It is also important to note that for the strong field cases with $M_L=0$, M_S and M have no quantum significance,[30a] but that each level with $M_L=0$ has a statistical weight $2S+1$, equal to the number of M_S values which would exist if there were a magnetic field parallel to the electric axis (the possible M_S values for this hypothetical case are indicated in brackets in Fig. 10). The diagrams are drawn, making use of Eqs. (22) and (27), for the case of so-called normal multiplets ($A>0$ in Eqs. (22) and (27).[34] The various energy intervals shown in Fig. 10 are, in sign and in relative magnitude, in qualitative agreement with what would be expected if the applied field were obtained by dividing the nucleus of an atom into two parts and separating these somewhat.

Quantum numbers for electron states of molecules (with fixed nuclei). Because of the fact that the energy in an axially symmetrical electric field depends on $|M_L|$ and $|M|$ rather than on M_L and M, and because of the importance of $|M_L|$ and $|M|$ in molecules, it is desirable to replace M_L, M_S, and M for the case of an atom in an electric field by new symbols. For a strong electric field we shall use Λ in place of $|M_L|$. At the same time we shall use Σ to represent $\pm M_S$, in such a way that when M_S and M_L have the same sign, Σ is positive, and that when they have opposite signs, Σ is negative. Also, we shall use $\Omega=|\Lambda+\Sigma|=|M|$. (If $\Lambda=0$, Σ and Ω do

[34] Cf. O. Stern, Zeits. f. Physik **23**, 476 (1922), and W. Pauli, Handbuch der Physik, Vol. 23, p. 251–2. The corresponding diagrams for inverted multiplets,—$A<0$, resulting in reversal of the energy order of the components of each multiplet,—are easily constructed. For inverted multiplets, the energy $F(0)$ in the strong field case would also usually lie above the energy $F(1)$, the relations of Fig. 10 being reversed.

not exist as quantum numbers.)[30a] According to these definitions $\Lambda \geqq 0$ always, while Ω takes on values $\Lambda+S,\ \Lambda+S-1, \cdots |\Lambda-S|$. Usually $\Lambda+\Sigma \geqq 0$, but if $S>\Lambda$, $\Lambda+\Sigma$ has one or more negative values so that for a given value of Λ, Ω may take on the same value twice (cf. the case $\Lambda=1$, $S=1\frac{1}{2}$ in Table III and Fig. 10). In a weak electric field, Λ and Σ do not exist as quantum numbers, but $\Omega = |M|$ remains.

The way in which Λ, Σ, and Ω are related to M_L, M_S, and M is shown in Table III for the case $L=1$, $S=1\frac{1}{2}$($^4\Pi$ state, cf. next paragraph and Fig. 10).

TABLE III.

M_L	M_S	M	Λ	Σ	$\Lambda+\Sigma$	Ω	State
$+1$ -1	$+1\frac{1}{2}$ $-1\frac{1}{2}$	$+2\frac{1}{2}$ $-2\frac{1}{2}$	1	$1\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$^4\Pi_{2\frac{1}{2}}$
$+1$ -1	$+\frac{1}{2}$ $-\frac{1}{2}$	$+1\frac{1}{2}$ $-1\frac{1}{2}$	1	$\frac{1}{2}$	$1\frac{1}{2}$	$1\frac{1}{2}$	$^4\Pi_{1\frac{1}{2}}$
$+1$ -1	$-\frac{1}{2}$ $+\frac{1}{2}$	$+\frac{1}{2}$ $-\frac{1}{2}$	1	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$^4\Pi_{\frac{1}{2}}$
$+1$ -1	$-1\frac{1}{2}$ $+1\frac{1}{2}$	$-\frac{1}{2}$ $+\frac{1}{2}$	1	$-1\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$^4\Pi_{-\frac{1}{2}}$

Molecular multiplets[28a] (*fixed nuclei*). Fig. 10 shows several interesting features which are of importance for the theory of molecular electronic states, since the latter (for fixed nuclei) are of the same nature as the states of atoms in a strong axially symmetrical electric field. First we may note how the strong field atomic levels group themselves in what we may reasonably classify as multiplets,[28a] and which we may correctly take as prototypes of the multiplets which are actually observed in molecular spectra. Each such multiplet has a definite Λ, which has a rôle very similar to that of L in an atomic multiplet; various different orientations of the spin S, each characterized by a definite Σ and giving a definite Ω, comprise the different sublevels of the multiplet; Ω plays a rôle similar to that of J in atomic multiplets. It is therefore natural to designate these multiplets in a manner similar to atomic multiplets, by symbols such as $^1\Pi$, $^2\Sigma$, $^3\Pi_2$, $^3\Pi_1$, $^3\Pi_0$, etc. In these the superscript gives the multiplicity $m(m=2S+1)$ exactly as for an atomic state, the main letter gives the value of Λ (with Σ, Π, Δ, Φ, Γ, $\cdots$ standing respectively for $\Lambda=0, 1, 2, 3, 4, \cdots$), and the subscript gives the value of $\Lambda+\Sigma$ (cf. Fig. 10 for examples). The value of $\Lambda+\Sigma$ rather than that of Ω is chosen for the subscript because it permits one to distinguish readily between physically different levels which have the same value of Ω, such as $^4\Pi_{\frac{1}{2}}$ and $^4\Pi_{-\frac{1}{2}}$ of Fig. 10 and Table III. When $\Lambda+\Sigma \geqq 0$, as usually, Ω is identical with $\Lambda+\Sigma$. For Σ levels we do not use a subscript, since Σ does not exist as a quantum number. For singlet levels subscripts are unnecessary, since Σ has only the one value 0.

Only the following types of electron levels have as yet been identified in practise; these have been found in the analysis of band spectra: $^1\Sigma$, $^2\Sigma$, $^3\Sigma$, $^1\Pi$, $^2\Pi$, $^3\Pi$, $^1\Delta$, $^2\Delta$, $^3\Delta$.

As should be clear from the discussion preceding Eq. (26), the behavior of the spin S in multiplets of the molecular type is very different than in atomic multiplets. As a consequence, the components of each such multiplet (F_0 and Λ fixed) are *equally spaced* (cf. Eq. (26) and definitions of Λ and Σ) according to the law

$$F-F_0=F(\Lambda)+A\Lambda\,\Sigma. \tag{27}$$

This relation is different from the Landé interval rule which holds for ordinary atomic multiplets (cf. Eq. (22)). Nevertheless the order of magnitude of the spacings of the component levels, determined by A of Eq. (22) or (27), is of the same order of magnitude in both cases. Examples of atomic and molecular multiplets may be compared in Fig. 10.

A consideration of the transition from strong to weak fields in Fig. 10 shows some points of interest. Although we may expect molecules to correspond pretty closely to the strong field case of Fig. 10, nevertheless this cannot be strictly true, and we may expect in practice to find some indications of a departure toward the weak field case. For the weak field case, each level with $\Omega>0$ has double statistical weight, because Ω represents $\pm M$; levels with $\Omega=0$, however, have unit statistical weight. For the strong field case, all levels with $\Lambda>0$ have double statistical weight (even if $\Omega=0$), because Λ represents $\pm M_L$. This double weight does not result [in the fixed center problem] in any doubling of the levels, except that in cases with $\Omega=0$, $\Lambda>0$, as $^3\Pi_0$, the level of double weight *does* split in the transition toward the weak field case (cf. Fig. 10).[30] In certain $^3\Pi_0$ levels of N_2 (second positive nitrogen bands) and C_2 (Swan bands), a very small splitting of this kind is actually found, indicating just a slight departure from the strong field case. In the corresponding $^3\Pi_1$ and $^3\Pi_2$ levels there is [for zero rotation], as expected, no splitting.

A somewhat similar state of affairs exists in certain Σ levels. In $^2\Sigma$ levels, no splitting occurs in the passage toward weak fields, but there is another interesting effect: Fig. 10 shows that although Ω is non-existent for the strict[30] strong-field case, an Ω should come into existence as soon as there is a departure toward weak fields.[35] In $^3\Sigma$ levels, passage toward weak fields gives a splitting into a level of double weight with $\Omega=1$ and a level of unit weight with $\Omega=0$; something similar happens with $^4\Sigma$ levels.[35] The $^3\Sigma$ levels of O_2 give evidence of a small splitting of the kind just described.

[35] It should be pointed out that the splittings just discussed for Σ levels depend on their being derived from atomic levels with $L>0$, like the P levels of Fig. 10. (Σ levels are derivable from every kind of atomic level, since for any L value one obtains Λ values equal to $L, L-1, \cdots, 0$.) For Σ levels derived from S levels, no splitting would occur [for zero rotation], since each S level would give just one Σ level, both the S and the Σ level being single, although of weight $2S+1$ (S=spin); also, Ω would be non-existent in such cases even for weak fields.

Selection rules and electron quantum numbers of molecules. The selection rules for an atom in a weak electric field are the same as those for an atom in a weak magnetic field. In a *strong* electric field the restriction on ΔJ is of course lost in proportion as J loses its significance as a quantum number. Just as in the case of strong magnetic fields, the rules $\Delta M_L = 0, \pm 1, \Delta M_S = 0$ should apply. ($\Delta M = 0, \pm 1$ also holds, but involves no restrictions beyond those already given by the rules for ΔM_L and ΔM_S.) In terms of Λ and Σ, these selection rules become $\Delta\Lambda = 0, \pm 1$ and $\Delta\Sigma = 0$ (also, $\Delta\Omega = 0, \pm 1$); but if $\Lambda = 0$ for one or both of two states between which a transition occurs, only $\Delta\Lambda = 0, \pm 1$ applies.[36] For the spin S we have the rule that intersystem transitions ($\Delta S > 0$) occur only with low intensity; for molecules with nuclei of small charge (H_2, He, He_2), $\Delta S = 0$ is almost a strict selection rule, but when a nucleus of high atomic number is present (e.g., Hg), transitions with $\Delta S = \pm 1$ are fairly intense. There are also other restrictions, especially in the case of symmetrical molecules (e.g., H_2) which will be considered later.

There is an important change in respect to l and L in strong electric fields: the selection rules $\Delta l = \pm 1$ and $\Delta L = 0, \pm 1$ lose their strictness, corresponding to the fact that a strong electric field has a tendency to break down the quantization of L and even of the l's.[32] In molecules J (in the sense in which it is used for atoms, as a resultant of L and S) and L are seldom of importance. Λ and S in molecules take the places which L and S have in atoms as the quantum numbers most characteristic of any multiplet electron level.

IIb. Theory of Molecular Energy Levels; Effects of Molecular Rotation

In Part I, we treated the total energy term F of a molecule as the sum of three parts, F^{el}, F^v, and F^r (cf. Eq. (5)), and determined how the spectrum depends upon the form of F'^v, F''^v, F'^r, and F''^r. The discussion was based on the assumption of a rotational energy of the simple form $F^r = B_v K(K+1) + \cdots$. Actually, this is not quite correct even in the simplest case $\Lambda = 0$, $S = 0$ ($^1\Sigma$ states); and in general, F^r has a variety of more complicated forms which we have now to consider. F^v, however, is fortunately of the same form $F^v = (v+1/2)\omega_e + \cdots$ for all types of electronic states, so that practically nothing new needs to be said about it in the following.

In Part II*a*, we have been considering the problem of the states of an atom in a strong axially symmetrical electric field. This problem is qualitatively the same as the problem of two fixed centers (cf. p. 85 above). In the transition to the real molecule we have, so far as nuclear vibration is concerned, only to add the proper expression for F^v to the F^{el} expression of the two-center problem. To take into consideration the rotation of the molecule, however, we have in general to take account of the fact that the latter often disturbs the characteristic angular momentum vectors and quan-

[36] The selection rules $\Delta\Lambda = 0, \pm 1$, $\Delta\Sigma = 0$ are formally more inclusive than $\Delta M_L = 0, \pm 1$, $\Delta M_S = 0$, since $\Lambda = \pm M_L$, and they apparently allow large changes in M_L (and in M_S and in M). There is, however, no real inconsistency, since levels $\pm M_L$ are indistinguishable.

tum numbers of the fixed-center problem. There is in general an interaction between the various electronic angular momenta and the nuclear angular momentum, so that we must consider F^{el} and F^{r}, and the associated quantum numbers, simultaneously. For sufficiently slow rotations, however, (*Hund's case a*), the quantum numbers Λ, Σ, and Ω of the fixed center problem are valid.

Before going into the problem of the interaction of rotation and electronic motions, a few remarks should be of interest in regard to the relative magnitude of the spacings of the component levels in a molecular multiplet as compared with the spacings of the associated vibrational and rotational levels. For most known band spectra, the spacings of the multiplet levels are much smaller than those of the vibrational levels, and often are comparable with those of the rotational levels. A typical example is that of the $^2\Pi$ normal state of NO. Here $A = 124$, $\Lambda = 1$, $\Sigma = \pm 1/2$ (cf. Eq. (27)), giving a doublet ($^2\Pi_{\frac{1}{2}}$ and $^2\Pi_{1\frac{1}{2}}$) of width $\Delta\nu = A\Lambda = 124$. The vibrational levels have a much larger spacing ($\omega_e = 1906$). Hence if one plots the energy levels, neglecting the rotation of the molecule, they appear as a series of vibrational levels each of which is double, and in band spectra involving this $^2\Pi$ state one finds double-headed bands. The most widely-spaced molecular multiplet yet known is a $^2\Pi$ state of HgH($\Delta\nu \sim 3700$). Here the spacing of the two components $^2\Pi_{\frac{1}{2}}$ and $^2\Pi_{1\frac{1}{2}}$ much exceeds the spacing of the vibrational levels ($\omega_e \sim 1980$), so that in the energy level diagram there appear to be two separate sets of vibrational levels, one for $^2\Pi_{\frac{1}{2}}$, one for $^2\Pi_{1\frac{1}{2}}$. Likewise in the spectrum, the bands associated with the $^2\Pi_{\frac{1}{2}}$ and $^2\Pi_{1\frac{1}{2}}$ levels appear like independent systems.

Singlet electron states: (Hund's case b').[36a] The simplest types of F^r functions and of band structure are those for *singlet* electronic states ($S = 0$). Here $\Sigma = 0$ and $\Lambda = \Omega$ always. So long as the effect of the nuclear rotation is not too great, Λ is a good quantum number and we have Hund's "case b'." Case b' is defined in general (cf. p. 108 below) as a case in which the energy of interaction of $\boldsymbol{S}^*$ with the rest of the molecule is practically zero, but in which the energy of orientation of $\boldsymbol{L}^*$ with reference to the electric axis is large enough so that Λ is a good quantum number. When $S = 0$, the first condition for case b' is automatically fulfilled.

The simplest relations exist in the case of $^1\Sigma$ states ($\Lambda = 0$, $S = 0$); these have been fully discussed in Part I. In such states the motion of the nuclei is practically one of pure rotation (but see below) with an angular momentum which we shall designate by $\boldsymbol{K}^* h/2\pi$. For $\Lambda > 0$, however, the motion is obviously gyroscopic, since we have both angular momentum of nuclear rotation (which we shall call $\boldsymbol{O}h/2\pi$)and electronic orbital angular momentum $\Lambda h/2\pi$. The two last-mentioned angular momentum vectors are perpendicular to each other and precess around their resultant $\boldsymbol{K}^* h/2\pi$, whose direction is fixed in space, since it represents the resultant angular momen-

[36a] Hund distinguished four important coupling cases which he called *a*, *b*, *c*, *d*; the names *b'* and *d'* given here are applied to certain variants of Hund's *b* and *d*; case *e* is a new case not discussed by Hund.

tum of the entire molecule. The precession frequency is the frequency of the nuclear rotation, since Λ is parallel to the line joining the nuclei and is therefore carried around with them in their rotation (cf. Fig. 11).

Now the resultant angular momentum P of any molecule is always subject to a quantum condition of the form $\int_0^{2\pi} P d\phi = 2\pi P$, $= k^* h$, where k is a quantum number which has either integral or half-integral values according as S is integral or half-integral. In the case now under discussion, we identify k^* with our K^*, and since $S=0=$integral, we conclude that K is a quantum number which has integral values. As we have already seen for the case of fixed nuclei, Λ is a quantum number whose possible values are the integers 0, 1, 2, $\cdots$.

For any definite electronic state, Λ is a constant, while O, which is not a quantum number, takes on various values, given by $(K^{*2}-\Lambda^2)^{1/2}$, such that K can have the values Λ, $\Lambda+1$, $\Lambda+2$, $\cdots$. For $\Lambda=0$, $O=K^*$. Fig. 11

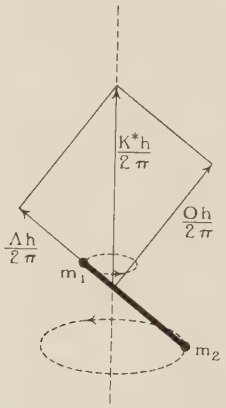


Fig. 11. Vector diagram and nuclear orbits (Hund's case b') for the case $\Lambda=2$, $K=3$, $S=0$ with two unequal nuclear masses ($m_1 > m_2$). The figure shows the vector diagram at a moment when the nuclei are in the plane of the paper. During the motion, $\boldsymbol{K}^*$ remains in the plane of the paper, while $\boldsymbol{\Lambda}$ always remains parallel and $\boldsymbol{O}$ perpendicular to the electric axis, i.e. to the line $m_1 \longleftrightarrow m_2$.

shows the motion of the nuclei, and the arrangement of the angular momentum vectors, for the case $\Lambda=2$, $K=3$. From the figure it is obvious that the minimum possible value of K is Λ; it may be remarked that, because of the fact that $K^*=[K(K+1)]^{1/2}$, $O>0$ even when $K=\Lambda$ (unless $\Lambda=0$).

The motion of the nuclei is really somewhat more complicated than has been represented in the preceding paragraphs. This is because of the fact that there is a component of electronic angular momentum, which we shall call $\boldsymbol{G}$, perpendicular to the electric axis.[37] $\boldsymbol{G}$ may be thought of as consisting primarily of the rapidly precessing component $\boldsymbol{L}_{\text{perp}}$ of $\boldsymbol{L}^*$ perpendicular to the electric axis, given by the vector equation $\boldsymbol{L}_{\text{perp}}=\boldsymbol{L}^*-\boldsymbol{\Lambda}$ (bold-face type will be

[37] J. H. Van Vleck, Phys. Rev. **31**, 600 (1928); J. H. Van Vleck and A. Frank, Proc. Nat. Acad. Sci. **15**, 539 (1929). Cf. also Hill and Van Vleck (reference 39, especially footnote 24), and Van Vleck (reference 43).

used throughout to indicate vectors) and by the scalar equation $L^2_{perp} = L^{*2} - \Lambda^2$. But strictly, $\boldsymbol{L}^*$ and so $\boldsymbol{L}_{\text{perp}}$ cannot in general be regarded as having even approximately a constant magnitude. In general, $\boldsymbol{G}$ then represents a quantity which varies rapidly in magnitude as well as in direction, corresponding to high frequency interchanges of angular momentum between the electron system and the nuclei,[32,37] but whose average magnitude is equal to that of $\boldsymbol{L}_{\text{perp}}$. The vector which we have called $\boldsymbol{O}$ includes $\boldsymbol{G}$ as a component, although when K is large, as is commonly true, G is small compared with O. $\boldsymbol{O}$ lies always in the plane determined by $\boldsymbol{K}^*$ and $\boldsymbol{\Lambda}$. $\boldsymbol{O}$ has to satisfy the vector equation $\boldsymbol{O} = \boldsymbol{N} + \boldsymbol{G}$, where we use $\boldsymbol{N}h/2\pi$ to designate the (instantaneous) angular momentum of the nuclei alone. $\boldsymbol{N}$ is a quantity whose magnitude varies very rapidly, but always in the neighborhood of its mean value, and whose direction also varies equally rapidly, generally within a moderate range, and in such a way as to keep $\boldsymbol{O}$ always in the proper plane and constant in magnitude. N is not a quantum number.

In Part I it was shown (cf. Eq. (12)) that the kinetic energy of rotation, plus potential energy of centrifugal expansion, of a non-vibrating molecule, is given by

$$E^r = P_\phi^2/2\mu r_e^2 - P_\phi^4/2\mu^2 a^* r_e^4 + \cdots$$

P_ϕ is the angular momentum of the nuclei alone. If in this equation we make the substitution $P_\phi = Nh/2\pi$, and also introduce B_e and D_e in accordance with Eqs. (3) and (4) of Part I, we get

$$F^r \equiv E^r/hc = B_e N^2 + D_e N^4 + \cdots$$

This is an accurate expression for the instantaneous energy of rotation even if, as in the case now under consideration, the nuclear angular momentum $Nh/2\pi$ is not a constant. For a vibrating molecule, this becomes (cf. Eqs. (3a), (4a), (5), (6))[38]

$$F^r = B_v N^2 + D_v N^4 + \cdots . \tag{28}$$

We wish now to re-express F^r in terms of quantum numbers. To do this, we first make use of the vector equation $\boldsymbol{N} = \boldsymbol{O} - \boldsymbol{G}$. Taking the scalar product of each side of this equation with itself, we get $N^2 = O^2 + G^2 - 2\,\boldsymbol{O}\cdot\boldsymbol{G}$, where $\boldsymbol{O}\cdot\boldsymbol{G}$ is the scalar product of $\boldsymbol{O}$ and $\boldsymbol{G}$. Since N varies rapidly during each rotation, we need in the F^r equation the mean values $\overline{N^2}$ of N^2 and $\overline{N^4}$ of N^4. We have $\overline{N^2} = O^2 + \overline{G^2} + \phi(K)$ where $\phi(K) \equiv -2\overline{\boldsymbol{O}\cdot\boldsymbol{G}} = -2O\rho$, if ρ is the mean value of the projection of $\boldsymbol{G}$ on $\boldsymbol{O}$.[38a] (It should be noted that O is here constant.) For a given $|\boldsymbol{G}|$, the projection of $\boldsymbol{G}$ on $\boldsymbol{O}$ is obviously equally likely to be positive or negative, except in so far as the motion of $\boldsymbol{G}$ is disturbed by the rotation of the molecule; hence we expect $\rho = 0$ for a rotationless molecule. When the molecule rotates, however, the

[38] Strictly speaking, there is a continual interchange of energy between vibration and rotation. The expression $E^r/hc = B_v N^2 - D_v N^4 + \cdots$ corresponds to the rotational energy *averaged over a complete vibration* (cf. Birge, Molecular Spectra in Gases, p. 112).

[38a] As used here, ρ has of course an entirely different meaning than in Part I.

symmetry of the motion of the vector G is disturbed so that we may expect a ρ, positive or negative, to develop. If the rotation is not too fast, we may reasonably assume $\rho \sim \delta O/2$, where δ is a small constant; then $\phi(K) \sim -\delta O^2$. For the present, however, we shall use the expression $\phi(K)$ without specifying its form.—For $\overline{N^4}$ one readily obtains an expression $\overline{N^4} = O^4 +$ terms of negligible importance.

Now substituting $O^2 = K^{*2} - \Lambda^2$ (cf. Fig. 11), and $K^{*2} = K(K+1)$, we get $\overline{N^2} = K(K+1) - \Lambda^2 + \overline{G^2} + \phi(K)$, and $\overline{N^4} = K^2(K+1)^2 + \cdots$. Substituting in the F^r expression, and adding F^{el} and F^v, we get [38]

$$F = F^{el} + F^v + B_v[K(K+1) - \Lambda^2 + \overline{G^2}] + \phi_i(K) + D_v[K^2(K+1)^2 + \cdots] + \cdots \quad (29)$$

F^v has the same meaning as previously (cf. Part I, Eq. (5)). The small terms given by $B_v(\overline{G^2} - \Lambda^2)$ are of no practical importance, since for any electronic state and v value, they enter as additive constants, which are not separable, in the analysis of band spectra, from the large terms $F^{el} + F^v$.

The "uncoupling term" $\phi(K)$ in Eq. (29) has been written $\phi_i(K)$, because if $\Lambda > 0$ it is double valued [$\phi_a(K)$ and $\phi_b(K)$], as discussed in the next paragraph. This term often takes the form (cf. second preceding paragraph) $\phi_i(K) = -\delta O^2 = -\delta[K(K+1) - \Lambda^2]$, so that the coefficient of $K(K+1)$ in Eq. (29) becomes $B_v - \delta = B_v{}^x$ instead of B_v; usually $\delta \ll B_v$, but sometimes δ is large enough so that it needs to be considered if one wishes to determine B_v, B_e, and r_e accurately. Sometimes $\phi_i(K)$ takes the form $\phi_i(K) \sim \epsilon K$, corresponding to a constant ρ (cf. second preceding paragraph), or $\phi_i(K) \sim \epsilon K - \delta K(K+1) + \cdots$; when such a term is added to $B_v K(K+1)$, the sum may be expressed in the form $B_v{}^x K^x(K^x+1)$, where $B_v{}^x$ and K^x differ slightly from B_v and K. Thus one gets apparent K values (K^x) differing slightly from integers. Slight deviations of this sort, and probably from this cause, are very commonly found in the analysis of band spectra.[12]

From the discussion on the states of an atom in a strong electric field (cf. pp. 91–3) it may be recalled that each state with $\Lambda > 0$ has a double statistical weight, because Λ really represents two states $\pm M_L$ which happen to have the same energy. In molecules, however, the rotation disturbs this exact coincidence of energies, so that a small splitting occurs,[39] each rotational level becoming a (usually narrow) doublet, whose components may be classified as sub-levels a and b. (In practise, the decision as to which doublet component shall be called a and which b is essentially arbitrary (cf. Part III), i. e. there is no one to one correspondence between the respective letters a and b and any theoretically specifiable properties of the two sub-levels.) The energies involved are given by the two values $\phi_a(K)$ and $\phi_b(K)$ of the small function $\phi_i(K)$. For $\Lambda = 0$, $\phi(K)$ has only a single value. When $\Lambda = 1$, each $\phi_i(K)$, provided the influence of the rotation is not too large, is given by $\phi_i(K) = -\delta_i K(K+1)$; hence the sub-levels a and b differ in respect to the apparent value of $B_v{}^x$ ($B_{v,a}{}^x = B_v - \delta_a$; $B_{v,b}{}^x = B_v - \delta_b$). This case is illustrated

[39] Cf. R. S. Mulliken, Phys. Rev. **28**, 1202 (1926); E. Hulthén, Zeits. f. Physik **46**, 349 (1927); R. de L. Kronig, Zeits. f. Physik **46**, 814 (1927); **50**, 347 (1928); E. L. Hill and J. H. Van Vleck, Phys. Rev. **32**, 267–272 (1928); J H Van Vleck, Phys. Rev. **33**, 484–489 (1929).

in the $^1\Pi$ levels in Fig. 13. For $\Lambda > 1$ the splitting $\phi_a(K) - \phi_b(K)$ is very small.[39] No quantum numbers can be assigned to the sub-levels a and b, but it is important to emphasize that they do *not* correspond to $+M_L$ and $-M_L$; *each* is really a mixture of $+M_L$ and $-M_L$ in a way which can only be understood and explained in terms of the new quantum mechanics.[39] This kind of doubling of rotational levels we shall designate as Λ-type doubling (formerly called σ-type doubling), because it occurs (for any value of S) in all rotational states where $\Lambda > 0$.

Singlet electron states: Hund's case d'. For the limiting case of fixed nuclei the "uncoupling term" $\phi(K)$ in Eq. (29) vanishes, and with it the Λ-type doubling,—except when $\Lambda = -\Sigma \neq 0$,—cf. $^3\Pi_0$ in Fig. 10. In singlet states of ordinary molecules, the relations which hold for fixed nuclei are only slightly modified by the nuclear rotation, giving small terms $\phi_s(K)$ and a narrow Λ-type doubling (case b'). The nature of the term $\phi(K)$, and the reason why it is double-valued when $\Lambda > 0$, can be better understood by a comparison of case b' with a condition (case d')[36a] which is close to the opposite limiting case in which the effect of the nuclear rotation is all-important.

For simplicity let us consider the case of a molecule composed of a core with $\Lambda = 0$ and a single outer electron, and let us consider as an example the case in which this outer electron has $l = 1$ and also has some definite value of the principal quantum number n. Then for the molecule in case b' we have a $^1\Sigma$ and a $^1\Pi$ state ($\Lambda = 0$ and 1). If the effect of the electric axis is large, I^* precesses very rapidly, the $^1\Sigma$ and $^1\Pi$ states are far apart in energy, each with its own set of (vibrational and) rotational energy levels. Now suppose we consider a series of orbits of increasing principal quantum number n, but all having $l = 1$. As the orbit gets larger, the influence of the electric axis on its energy rapidly gets less; this decreasing relative importance of the electric axis is of the same nature as if, with a *fixed* orbit size, the nuclei were gradually brought together. The rate of precession of I^* and the energy interval beween $^1\Sigma$ and $^1\Pi$ diminish rapidly as the size of the orbit increases. Finally, the rate at which I^* tends to precess around the electric axis becomes small compared with the rate at which the latter is carried around in space by the rotation of the nuclei. The precession of I^* around the electric axis breaks down, and Λ ceases to have any significance as a quantum number.

From the point of view of the electron in its large orbit, the two nuclei are now close together and (with the inner electrons) form a whirling mass whose only distinguishable axis is the axis of rotation of the nuclei (perpendicular to the electric axis). The nuclei rotate (and vibrate) almost as if the outer electron had been removed completely, their rotation being characterized in first approximation by a rotational quantum number R whose possible values are 0, 1, 2, $\cdots$, and by a rotational energy (cf. below, following Eq. (30))

$$F^r = B_v[R(R+1) + \overline{G^2}] + D_v[R^2(R+1)^2 + \cdots] + \cdots$$

[39a] The designations a and b as used for sub-levels have no connection with the same designations in Hund's cases a and b.

At the same time the outer electron moves in its orbit in almost the same way as if the nuclei were united. The total energy does, however, usually depend somewhat on the orientation of the plane of rotation of the electron with reference to that of the nuclei; in other words, it depends slightly on the angle between $\boldsymbol{R}^*$ and $\boldsymbol{l}^*$. Hence $\boldsymbol{R}^*$ and $\boldsymbol{l}^*$ give a quantized resultant $\boldsymbol{K}^*$, which, since it represents the resultant angular momentum of the molecule, is fixed in space. K has the values $R+l, R+l-1, \cdots |R-l|$. $\boldsymbol{R}^*$ and $\boldsymbol{l}^*$, exerting a torque on each other, slowly precess about $\boldsymbol{K}^*$. The relations just

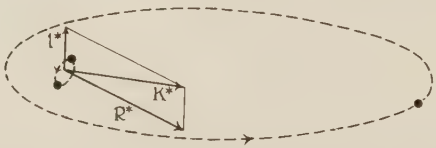


Fig. 12. Vector and orbit diagram for Hund's case d', with $S=0$. The large and the small ellipse respectively represent the orbit of the outer electron and of the nuclei. The vectors $\boldsymbol{l}^*$ and $\boldsymbol{R}^*$, which are intended to be respectively perpendicular to the planes of the large and small ellipses, represent the electronic and nuclear angular momenta. $\boldsymbol{R}^*$ and $\boldsymbol{l}^*$, and with them the corresponding orbit planes, are supposed to be precessing slowly around their resultant $\boldsymbol{K}^*$, which has a fixed direction and magnitude. The figure is drawn for the case of two equal nuclei ($m_1=m_2$), and is directly applicable to the high-quantum singlet states of H_2 and He_2.

described are depicted in Fig. 12. Examples of this situation, (*case d′ of Hund*[24]) are found in the higher-quantum singlet [and triplet] electron states of H_2 and He_2.[40] The total energy for Hund's case d' is

$$F=F^{el}+F^{v}+B_v[R(R+1)+\overline{G^2}]+f(R,K-R)+D_v[R^2(R+1)^2+\cdots]+\cdots \quad (30)$$

Here the "coupling term" $f(R, K-R)$ is a small term depending on R and on the relative orientation of $\boldsymbol{l}^*$ and $\boldsymbol{R}^*$. The term $B_v\overline{G^2}$ is needed to take account of the varying perpendicular component of orbital angular momentum of the inner electrons. In general, the inner electrons might also have a Λ, giving $B_v[R(R+1)-\Lambda^2+\overline{G^2}]$ in Eq. (30).

The fine structure, corresponding to $f(R, K-R)$ in Eq. (30), in the rotational levels in Hund's case d', is all that remains of the division into states with differents Λ values which exists in Hund's case b'. In the limiting case where the coupling of $\boldsymbol{l}^*$ to the electric axis is negligible, even this would disappear and there would be, for each value of l, a single set of rotational levels determined by R alone. Figs. 13 and 14 show, for $l=1$ and $l=2$ respectively, how the energy levels for case b' and case d' are related.[41] The method of correlation is based on two rules analogous to those used in connection with Figs. 9 and 10. These are, (1) K, representing total angular

[40] W. Weizel, He_2 bands: Zeits. f. Physik **52**, 175 (1928); **54**, 329 (1929); H_2 bands, Zeits. f. Physik **55**, 483 (1929). He_2 bands, review by W. E. Curtis, in the discussion on Molecular Spectra and Molecular Structure in the Faraday Society Transactions (Sept., 1929).

[41] For a discussion of the energy levels in Hund's case d and their relation to Λ-type doubling through the transition to case b, cf. Hill and Van Vleck, reference 39, pp. 267–272.

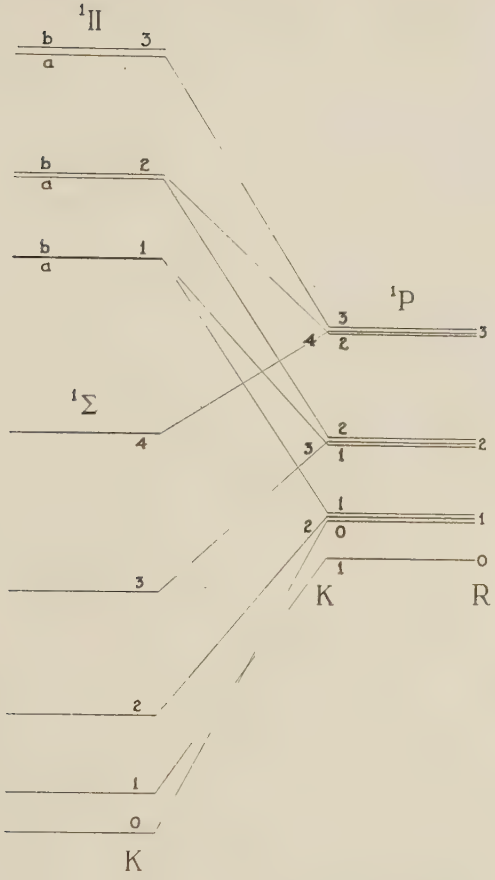


Fig. 13. Relation between case b' and case d' energy levels for $L=l=1$, $S=0$. In case b' we have $\Lambda=0$ ($^1\Sigma$ levels) and $\Lambda=1$ ($^1\Pi$ levels), with $K=\Lambda, \Lambda+1, \cdots$, and with Λ-type doublet in the $^1\Pi$ state; the doublet widths have been assumed proportional to $K(K+1)$, in agreement with both theory and experiment.[43] The $^1\Sigma$ and $^1\Pi$ sets of levels should be imagined as much more widely separated than in the figure. In case d' we have $R=0, 1, \cdots, K=R, R\pm1$; the positions of the three levels $K=R, R\pm1$ have been calculated[41] in accordance with certain equations of Hill and Van Vleck (Phys. Rev. **32**, 269, 1928, Eqs. (42) with $k=1$) which are based on the reasonable assumption that the energy of interaction of $\boldsymbol{R}^*$ and $\boldsymbol{l}^*$ is given by $(A\overline{l^2}_{par}+\text{const.})$, where $\overline{l^2}_{par}$ is the average value of the square of the projection of $\boldsymbol{l}^*$ on the electric axis; the levels have been calculated for the case $A=B_v/2$, i.e. $A=B/2$ was substituted in Hill and Van Vleck's formula before making the calculation. The case b' levels (here $\overline{l^2}_{par}=\Lambda^2$) correspond to $A\gg B_v$. If A were negative, as it might be, the order of the $^1\Sigma$ and $^1\Pi$ levels would be reversed in case b', and the correlations with case d' would be different. In Fig. 13, except for the level $K=0$, the correlations may be expressed as follows: the $^1\Sigma$ levels of case b' go over into levels with $K=R+1$ in case d', the $^1\Pi_a$ levels into levels with $K=R$, and the $^1\Pi_b$ levels into levels with $K=R-1$. (It should be remarked that the designations a and b in Fig. 13 are, as always in the last analysis, arbitrary.)

momentum, remains constant for any level; (2) the members of any group of levels having the same K value never cross one another.—For $l=0$, cases b' and d' would be identical; for $l>2$, correlations similar to those of Figs. 13 and 14 would hold.

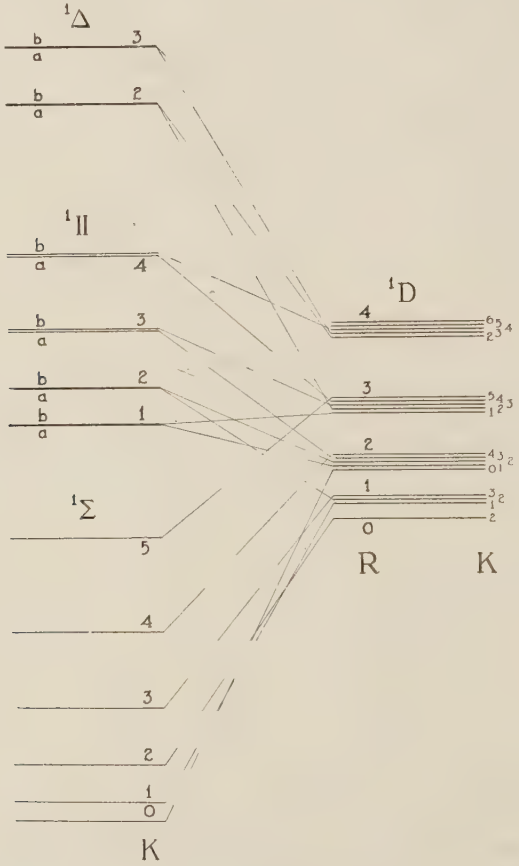


Fig. 14. Relation between case b' and case d' energy levels for $l=2$, $S=0$. This is similar to Fig. 13, except that the uniform spacings of the levels shown here, for fixed R and variable K in case d', have no significance; Hill and Van Vleck give no theoretical formula for the case $l=2$. In agreement with theory and experiment,[43] the Λ-type doublet widths are shown much narrower for the Δ than for the Π levels. If the order of the energy levels $^1\Sigma$, $^1\Pi$, and $^1\Delta$ were inverted in case b', as is possible, the correlations with case d' would be different. In Fig. 14, except for the levels with $K=0$ and 1, the correlation is as follows: $^1\Sigma$ goes into $K=R+2$, $^1\Pi_a$ into $K=R+1$, $^1\Pi_b$ into $K=R$, $^1\Delta_a$ into $K=R-1$, and $^1\Delta_b$ into $K=R-2$. (As in Fig. 13, the designations a and b are essentially arbitrary.) Just such a correlation as this appears actually to exist in the states of the He_2 molecule[40] having $l=2$ for the series electron.

Figs. 13 and 14 show how it is that in Λ-type doubling the a and b sublevels, which would exactly coincide ($\phi_i(K)=0$) for the limiting case of fixed nuclei, must needs separate more and more widely as the disturbance of Λ by the molecular rotation increases,[41] in order to enter into new groupings in case d'.

Multiplet electron states: Hund's case a. In multiplet electronic states ($S>0$) the presence of a spin gives rise to two important coupling cases, Hund's cases *a* and *b*;[24] these both degenerate for singlet states into what we have called case *b'*. In *Hund's case a*[24] the spin S is quantized with reference to the electric axis, because of the magnetic field along this axis produced by Λ. This quantization gives Σ, as discussed in the previous section on the atom in a strong electric field. Case *a* is possible only if $\Lambda>0$, for reasons previously discussed (cf. ref. 30a). Next we have $\Omega=|\Lambda+\Sigma|$; Ω and $\boldsymbol{O}$ then form a resultant J^* exactly as Λ and $\boldsymbol{O}$ formed a resultant $\boldsymbol{K}^*$ in the case of singlet states (cf. Fig. 11 and accompanying discussion).[41a] J is a quantum number, since it corresponds to the resultant angular momentum of the molecule. It has values Ω, $\Omega+1$, $\Omega+2$, $\cdots$, which are integral or half-integral according as S(and therefore Σ and Ω) are integral or half-integral. (The molecular J is not the same as J for an *atom* except in the fact that each represents a total angular momentum). In complete analogy to Eq. (29), except for the addition of the energy of interaction of Λ and Σ (cf. Eq. (27)) and the occurrence of a term in $S_{perp}{}^2$ ($S_{perp}=S^{*2}-\Sigma^2$),[42] the energy is given by

$$F=F_0{}^{el}+F^v+A\Lambda\Sigma+B_v[J(J+1)-\Omega^2+\overline{G^2}+S_{perp}{}^2]+\phi_i(\Sigma,J)+D_vJ^2(J+1)^2+\cdots \quad (31)$$

$F_0{}^{el}+A\Lambda\Sigma$ may be considered to represent electronic energy, different values of Σ, for fixed $F_0{}^{el}$ and Λ, giving different components of a molecular multiplet. For each such value of Σ there is a set of vibrational and rotational levels. The rotational levels in each set are all double (Λ-type doubling), with two values of $\phi_i(\Sigma, J)$ for each value of Σ and J [$\phi_a(\Sigma, J)$ and $\phi_b(\Sigma, J)$]. The function F^v is found empirically to be practically independent of the value of Σ in a multiplet unless A is very large. Similarly, the coefficients B_v (and presumably also D_v) are almost independent of Σ. (When the conditions of case *a* are not well fulfilled there is, however, an apparent dependence of B_v on Σ,—(cf. (Eq. 46)). The small functions ϕ_i (Σ, J), unlike the B_v terms, differ greatly according to the value of Σ, the more so the larger A is.[43]

Multiplet states: case b. In Hund's case *a*, we think of $\boldsymbol{S}^*$ as precessing around the electric axis with a frequency $c\omega_S=c\left|\partial F/\partial\Sigma\right|=cA\Lambda$ (cf. Eq. (23)), making with that axis a constant angle θ such that $\cos\theta=\Sigma/S^*$. The conditions necessary for Hund's case *a* are well fulfilled so long as

[41a] But in Hund's case *a* $\boldsymbol{O}$ includes $\boldsymbol{S}_{perp}$ as well as $\boldsymbol{G}$ as a component (cf. Ref. 42).

[42] Cf. Hill and Van Vleck, reference 39. If $\boldsymbol{N}h/2\pi$ is the nuclear angular momentum, we have the vector equation $\boldsymbol{N}=\boldsymbol{O}-\boldsymbol{G}-\boldsymbol{S}_{perp}$. Taking the scalar product of each side with itself, $N^2=O^2+G^2+S^2{}_{perp}-2\boldsymbol{O}\cdot\boldsymbol{G}-2\boldsymbol{O}\cdot\boldsymbol{S}_{perp}+2\boldsymbol{G}\cdot\boldsymbol{S}_{perp}$. Each of the last two terms has zero as its average value. Therefore for the average value $\overline{N^2}$, using also the relation $O^2=J^{*2}-\Omega^2$, we have $\overline{N^2}=J(J+1)-\Omega^2+\overline{G^2}+\overline{S^2}_{perp}-2\,\overline{\boldsymbol{O}\cdot\boldsymbol{G}}$. Now we write $\overline{S^2}_{perp}$ as $S^2{}_{perp}$, since the latter is constant. Proceeding as in the case of Eq. (28), and putting $-2B_v\overline{\boldsymbol{O}\cdot\boldsymbol{G}}=\phi_i(\Sigma, J)$, we get Eq. (31).

[43] Theory, J. H. Van Vleck, Phys. Rev. **33**, 467 (1929); compilation of empirical data, R. S. Mulliken, Phys. Rev. **33**, 507 (1929).

$\omega_S \gg \omega_J$, where $c\omega_J$ is the frequency of rotation of the nuclei and therefore of the electric axis. [$c\omega_J$ is obtained by taking $c\left|\partial F/\partial J\right|$ in Eq. (31): $c\omega_J \sim cB_v(2J+1)$.] But if B_v or J is sufficiently large or A sufficiently small, so that $\omega_S \ll$ (the calculated) ω_J, then the precession of $\boldsymbol{S}^*$ breaks down, and Σ ceases to exist as a quantum number. [This is similar to the failure of Λ in Hund's case d',—cf. discussion preceding Eq. (30).] Λ, however, still remains quantized, and Λ and $\boldsymbol{O}$ now form a resultant $\boldsymbol{K}^*$, around which they precess, in exactly the same manner as for case b' with $S=0$ (cf. Fig. 10, Eq. (29), and accompanying discussion). For the spin, which is here unable to keep up with the (relatively) very rapid precession of Λ, the only axis now distinguishable is that of $\boldsymbol{K}^*$. Parallel to this axis there is usually a small magnetic field (in regard to the cause of this, cf. p. 107 below), and $\boldsymbol{S}^*$ therefore precesses around $\boldsymbol{K}^*$. The resultant $\boldsymbol{J}^*$ of $\boldsymbol{K}^*$ and $\boldsymbol{S}^*$ is a vector fixed in space, since $J^*h/2\pi$ is the resultant angular momentum of the molecule. J is a quantum number, whose possible values are given by $J=K+S, K+S-1, \cdots \left|K-S\right|$. K, S, and J here are respectively closely analogous to L, S, and J in an atom (cf. discussion preceding Eq. (22)). For each K value, the group of energy levels associated with it, corresponding to different J values, forms a sort of tiny multiplet. The state of affairs just described is typical of *Hund's case b*.

The energy equation for Hund's case b in multiplet states can be readily obtained by a slight generalization of the Eq. (29) which holds for case b' singlet states. It is only necessary to add to Eq. (29) a small term $f(K, J-K)$ to take account of the energy of interaction of $\boldsymbol{K}^*$ and $\boldsymbol{S}^*$, and to replace $\phi_i(K)$ of Eq. (29) by a similar function $\phi_i(K, J)$ which, for any specified K and J, is double-valued [$\phi_a(K, J)$ and $\phi_b(K, J)$] except for $\Lambda=0$, where it is single-valued. The values of $\phi_i(K, J)$, where i is either a or b, in general differ slightly (but only slightly,[43] in marked contrast to the large dependence of $\phi_i(\Sigma, J)$ in case a on the value of Σ) for the $(2S+1)$ different J values associated with a given K value. Thus we have

$$F=F_0{}^{e1}+F^v+B_v[K(K+1)-\Lambda^2+\overline{G^2}] \\ +f(K,J-K)+\phi_i(K,J)+D_vK^2(K+1)^2+\cdots \quad (32)$$

Since for any value of K there are in general $2S+1$ components in the spin fine structure, there are in general altogether either $2S+1$ or $2(2S+1)$ fine structure components for each value of K, according as $\Lambda=0$ or $\Lambda>0$. (When $K<S$, however, there are fewer components.) The relative scales of the two types of fine structure which exist when $\Lambda>0$ differ from one molecule and electron state to another; sometimes the Λ-type doubling (given by $\phi_a-\phi_b$) is coarser than the spin fine structure, sometimes the reverse.

The small term $f(K, J-K)$ in Eq. (32) consists in general of several parts. (1) First we have the energy of interaction of $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ which is given, as in Eqs. (22) and (26), by $AL^*S^*\cos(\boldsymbol{L}^*, \boldsymbol{S}^*)$. This can be written in the form $A\boldsymbol{L}^*\cdot\boldsymbol{S}^*$, where $\boldsymbol{L}^*\cdot\boldsymbol{S}^*$ is the scalar product of the vectors $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$. By taking components of $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ along $\boldsymbol{K}^*$, this expression reduces to $A\Lambda^2(J^{*2}-K^{*2}$

$-S^{*2})/2K^{*2}$.[44] This drops out if $\Lambda=0$. (2) Next we have a small term resulting from the interaction of $\boldsymbol{S}^*$ with the small magnetic field, parallel to $\boldsymbol{K}^*$, which is developed by the rotation of the molecule. This field is that resulting from the forced rotation of the whole electron system of the molecule around $\boldsymbol{K}^*$, modified (but probably usually only to a slight extent) by the field of the rotating nuclei themselves. Assuming that the field is parallel and proportional[44a] to $\boldsymbol{K}^*$, the energy of interaction of $\boldsymbol{S}^*$ with it is given by an expression of the form $\gamma(\boldsymbol{K}^*\cdot\boldsymbol{S}^*)$. This reduces to $((\gamma/2)\,(J^{*2}-K^{*2}-S^{*2})$. This expression should hold provided K is not too large.[45] (3) Finally, as Kramers has shown,[46] there is for $S>1/2$ an energy of interaction $w(K, J\text{-}K)$ of the individual spins which make up the resultant S. For Σ states at least, this energy is proportional to $3\overline{\cos^2\theta}-1$, where θ is the angle between $\boldsymbol{S}^*$ and the electric axis. For $^3\Sigma$ states, where there are two individual spins, Kramers finds

$$w(K,+1)=-\epsilon[1-3/(2K+3)]\,;\,w(K,0)=+2\epsilon\,;\,w(K,-1)=-\epsilon[1+3/(2K-1)];$$

the states with $J=K\pm1$ form a narrow doublet whose center is separated from the state $J=K$ by an approximately constant energy interval 3ϵ. (The energy $\gamma\boldsymbol{K}^*\cdot\boldsymbol{S}^*$ must also be added, but usually $\gamma<\epsilon$). This case is found experimentally in the $^3\Sigma$ normal state of O_2.[46] (It may be recalled that just such a splitting of $^3\Sigma$ levels into two components, independently of any rotation of the molecule, was predicted above by a consideration of the states of a ^3P atom in an electric field: cf. Fig. 10 and p. 95).

From the above, we have

$$f(K,J-K)=(\tfrac{1}{2})\left\{[A\Lambda^2/K(K+1)]+\gamma\right\}\left\{J(J+1)-K(K+1)-S(S+1)\right\}+w(K,J-K). \quad (33)$$

The last term vanishes (or at least becomes independent of $J-K$) unless $S>1/2$.

[44] $\boldsymbol{L}^*$ is composed of $\boldsymbol{\Lambda}+\boldsymbol{L}_{perp}$, but $\overline{\boldsymbol{L}_{perp}\cdot\boldsymbol{S}^*}=0$; $\boldsymbol{\Lambda}$ has a component Λ^2/K^* parallel to $\boldsymbol{K}^*$ (the component perpendicular to $\boldsymbol{K}^*$ gives $\overline{\boldsymbol{\Lambda}_{perpk}\cdot\boldsymbol{S}^*}=0$). (The result is the same even if we recognize that $\boldsymbol{L}^*$ is not a constant, since in *any* case $\overline{\boldsymbol{G}\cdot\boldsymbol{S}^*}\sim0$, if $\boldsymbol{G}$ represents the component of electronic orbital angular momentum perpendicular to the electric axis; thus only the parallel component $\boldsymbol{\Lambda}$ needs to be considered.) $\boldsymbol{S}^*$ has a component $(J^{*2}-K^{*2}-S^{*2})/2K^*$ parallel to $\boldsymbol{K}^*$.

[44a] Assuming (*a*) that the field is proportional to the component ρ (cf. pp. 99–100) of electronic orbital angular momentum parallel to $\boldsymbol{O}$ developed by the rotation, and (*b*) that ρ is proportional to O for not too large K values (cf. p. 100), we should have, for not too large K values, a field parallel and proportional to O. This may be replaced without serious error by the assumption of a field parallel and proportional to $\boldsymbol{K}^*$ since, except for the smallest K values where the term $\gamma(\boldsymbol{K}^*\cdot\boldsymbol{S}^*)$ is of negligible importance anyway, $\boldsymbol{O}$ and $\boldsymbol{K}^*$ are practically the same. The results obtained from these assumptions are in harmony with the known experimenta data, at least for Σ states (where $\boldsymbol{O}\equiv\boldsymbol{K}^*$).

[45] Cf. E. C. Kemble, Molecular Spectra in Gases, pp. 345–7; R. S. Mulliken, Phys. Rev. **30**, 149 (1927); F. Hund, Zeits. f. Physik **42**, 96 (1927); E. Hulthén, Zeits. f. Physik **50**, 319 (1928); and especially J. H. Van Vleck, Phys. Rev. **33**, 498–500 (1929).

[46] H. A. Kramers, Zeits. f. Physik **53**, 422 (1929). Interpretation of O_2 bands, R. S. Mulliken, Phys. Rev. **32**, 880 (1928).

In the important special case of case *b doublet* spectra ($S=1/2$), we have $J=K\pm 1/2$. Substitution in Eq. (33) with the $w(K, J-K)$ term omitted gives

$$f(K,-\tfrac{1}{2})=-A\Lambda^2/2K-\gamma(K+1)/2=-A\Lambda^2/(2J+1)-\gamma(2J+3)/4 \quad (34)$$

$$f(K,+\tfrac{1}{2})=A\Lambda^2/2(K+1)+\gamma K/2=A\Lambda^2/(2J+1)+\gamma(2J-1)/4. \quad (35)$$

The doublet separation, $f(K, +1/2)-f(K, -1/2)$, associated with any value of K is then

$$\Delta f(K)=[A\Lambda^2/K(K+1)+\gamma](K+\tfrac{1}{2}). \quad (36)$$

Multiplet states: *case b'*. It frequently happens that the spin fine structure given by $f(K,J-K)$ is extremely narrow or even quite undetectable. This often occurs in $^2\Sigma$ states, especially for small K values. It is also typical of the triplet states ($S=1$) of H_2 and He_2, where A and γ are very small because of the small charges of the nuclei. (As is well known from theory and from experimental data on atomic spectra, especially for H and He, A is always very small for the lightest elements; while, as Van Vleck has shown,[45] the magnitude of γ is proportional to that of A). When $f(K,J-K)\sim 0$, Eq. (32) becomes practically identical with the equation for singlet states (Eq. (29)) and there is no way of telling empirically whether S is zero or greater; K can then be determined, but not J. In such a case the precession of $\mathbf{S}^*$ about $\mathbf{K}^*$ is so slow (frequency $=c\{\partial/\partial J[f(K, J-K)]\}$) that $\mathbf{S}^*$ may be regarded as practically free, while J becomes practically meaningless as a quantum number. It seems desirable to distinguish this situation as *case b'*.[36a] Case *b'* is merely a form of case *b* in which the interaction of S^* with the rest of the molecule is negligibly small. This condition is automatically fulfilled when $S=0$, as already noted in the discussion of singlet states.

Multiplet states: *cases d and d'*. In spite of the absence of *spin* fine structure, the case *b'* states of H_2 and He_2 often show a rather coarse Λ-type doubling ($\phi_a(K)-\phi_b(K)$ large). This increases with increase in the principal quantum number of the orbit of the outer electron, until finally we pass for very large orbits to *Hund's case d'* as described above in connection with Eq. (30). This happens for both singlet and triplet states ($S=0$ and 1), but there is no practical difference between these cases, since the spin, when present, is essentially free. If there *were* a slight interaction of S with the rest of the molecule, we should expect $\mathbf{S}^*$ to orient itself with reference to the resultant $\mathbf{K}^*$ of $\mathbf{L}^*$ and $\mathbf{R}^*$ to give a final resultant J^*. We call this last case *d*, and the case in which J does not exist as a quantum number, case *d'*. We return now to cases *a* and *b*.

Doublet states: *transition cases between a and b*. When $\Lambda=0$ (Σ states) or $S=0$ (singlet states), Hund's case *a* is impossible, and case *b* (or *b'*, *d*, or *d'*) is found; when $\Lambda>0$, either *a* or *b* is possible, but one also very frequently finds intermediate cases which do not closely approximate either *a* or *b*. Very often case *a* is approximated for the lowest J values, but case *b* for high J values. For $S=1/2$ (*doublet states*), Hill and Van Vleck[46] have obtained an

equation which gives the main term of F^r, plus the interaction term of Λ and S^*, exactly for the whole range of intermediate cases. This is

$$F=F_0{}^{el}+F^v+B_v\{(J+\tfrac{1}{2})^2-\Lambda^2\pm\tfrac{1}{2}[4(J+\tfrac{1}{2})^2 -4A\Lambda^2/B_v+A^2\Lambda^2/B_v{}^2]^{1/2}+\overline{G^2}\}+\phi_i(J)+\cdots \quad (37)$$

(In Eq. (37), no account has been taken of the energy which corresponds to the small term proportional to γ in Eq. (33). Allowance must therefore be made for this omission when Eq. (37) is applied.) The + and − signs in Eq. (37) give the energy values corresponding to the two orientations of the spin which exist in cases *a* and *b* and in all intermediate cases as well. Neglecting Λ-type doubling, which exists for all levels if $\Lambda>0$, there are thus just two energy levels for each value of J, except for the lowest value $J=\Lambda-1/2$, where there is only a single level. We shall now consider the forms which Eq. (37) takes for various values of A/B_v and of J. First we shall consider the special J value $J=\Lambda-1/2$; then we shall consider the expansion of Eq. (37) for large values of $|A|$, corresponding to case *a*, and after that we shall see what forms it takes for small values of A/B_v (case *b*).

For $J=\Lambda-1/2$ there is only one energy level (when Λ-type doubling is neglected), as can be seen by consideration of the limiting cases *a* and *b*. For example, in case *a* with $\Lambda=1$, the value $J=1/2$ occurs only in the $^2\Pi_{\frac{1}{2}}$ series of levels ($\Sigma=-\frac{1}{2}$, $\Omega=\frac{1}{2}$), and not in the $^2\Pi_{1\frac{1}{2}}$ series ($\Sigma=+\frac{1}{2}$, $\Omega=1\frac{1}{2}$), because of the restriction $J\geqq\Omega$; for all other J values, however, there is obviously one level in *each* of the two series. Corresponding results hold for *case b* with $\Lambda=1$: first, we have $K\geqq\Lambda$; then we have $J=K\pm\frac{1}{2}$; $K=\Lambda=1$ then yields $J=\frac{1}{2}$, $1\frac{1}{2}$, while $K=2$ yields $J=1\frac{1}{2}$, $2\frac{1}{2}$, and so on; thus $J=\frac{1}{2}$ occurs only once, but all other J values twice.—For $J=\Lambda-\frac{1}{2}$ the correct energy value is obtained from Eq. (37) by using the upper (+) sign when $-\infty<A/B_v<2$, and the lower (−) sign when $2<A/B_v<+\infty$. Eq. (37) then simplifies in both cases to the following form:

$$F(J=\Lambda-\tfrac{1}{2})=F^{el}+F^v-A\Lambda/2+B_v(\Lambda+\overline{G^2})+\phi_i(\Lambda-\tfrac{1}{2})+\cdots. \quad (38)$$

This expression is identical with that which is obtained for $J=\Lambda-\frac{1}{2}$ from the energy formulas for *either case a* (Eq. (31)) or case *b* (Eqs. (32) and (33), neglecting γ and w in Eq. (33)); this can readily be verified by making the proper substitutions for J, K, S, and Σ in Eqs. (31)–(33).

For large positive values of A (*normal* case *a*) expansion of Eq. (37) gives[47]

$$F=F_0{}^{el}+F^v\pm(\tfrac{1}{2})A\Lambda+B_v\{J(J+1)-(\Lambda\pm\tfrac{1}{2})^2 +(\tfrac{1}{2})+\overline{G^2}\pm[(J+\tfrac{1}{2})^2-\Lambda^2]B_v/A\Lambda+\cdots\}+\phi_i(J)+\cdots \quad (39)$$

For large negative values of A (*inverted* case *a*) expansion gives[47]

$$F=F_0{}^{el}+F^v\mp(\tfrac{1}{2})A\Lambda+B_v\{J(J+1)-(\Lambda\mp\tfrac{1}{2})^2 +(\tfrac{1}{2})+\overline{G^2}\mp[(J+\tfrac{1}{2})^2-\Lambda^2]B_v/A\Lambda+\cdots\}+\phi_i(J)+\cdots. \quad (40)$$

[47] E. L. Hill and J. H. Van Vleck, Phys. Rev. **32**, 261–2 (1928). Eq. (37) as given above differs from the equation given by Hill and Van Vleck in the substitution of B_v for B, and in the addition of the unimportant additive term $B_v\overline{G^2}$ which was dropped by them.

In Eqs. (39) and (40), the upper sign, in each case of a double sign, corresponds to the upper (+) sign in Eq. (37). Bearing in mind that Eqs. (39) and (40) apply only to doublet levels ($S=\frac{1}{2}$), comparison with the equation for case a (Eq. 31) shows that the quantities $\pm(\frac{1}{2})$ or $\mp(\frac{1}{2})$ in Eqs. (39) and (40),—*e.g.*, $\pm(\frac{1}{2})$ in $\pm(\frac{1}{2})A\Lambda$ and in $(\Lambda\pm\frac{1}{2})^2$ of Eq. (39),—are to be identified with Σ of Eq. (31), while $+(\frac{1}{2})B_v$ is identified as $B_v S^2_{perp}$ of Eq. (31).

It should be mentioned that equations for triplet states, analogous to Eqs. (39) and (40) for doublet states, are given by Hill and Van Vleck.[46]

The following expansion of Eq. (37) represents a good approximation for a range of values of A/B_v corresponding fairly well to case b and extending, roughly, from -2 to $+6$.

$$F=F_0{}^{el}+F^v+B_v\left[(J+\tfrac{1}{2})^2\pm(J+\tfrac{1}{2})-\Lambda^2+\overline{G^2}\right]$$
$$\mp\left[A\Lambda^2/2(J+\tfrac{1}{2})\right]\left[1-A/4B_v+A\Lambda^2/4B_v(J+\tfrac{1}{2})^2-A^2\Lambda^2/8B_v{}^2(J+\tfrac{1}{2})^2+\cdots\right]$$
$$+\phi_i(J)+(\gamma/2)\left[J(J+1)-K(K+1)-S(S+1)\right]+\cdots. \tag{41}$$

The last term in Eq. (41) does not come from the expansion, but has been added (cf. Eq. (33)) in accordance with a remark made above immediately after Eq. (37).

If we neglect the terms in A^2 and A^3, Eq. (41) reduces to the following form which agrees with the case b equations (Eqs. (32), (33)), provided, in the case of the upper and lower signs respectively in Eq. (41), we put $J=(K-\frac{1}{2})$ and $J=K+\frac{1}{2}$.

$$F=F_0{}^{el}+F^v+B_v\left[K(K+1)-\Lambda^2+\overline{G^2}\right]$$
$$+(A\Lambda^2/2)\left[\text{times}-1/K(\text{for } J=K-\tfrac{1}{2}),\ \text{or times}\ +1/(K+1)\right.$$
$$\left.(\text{for } J=K+\tfrac{1}{2})\right]+\phi_i(K,J)+(\gamma/2)\left[J(J+1)-K(K+1)-S(S+1)\right]+\cdots. \tag{42}$$

In order to see that this agrees with Eqs. (32, 33) it is necessary to complete the substitutions $J=K\mp\frac{1}{2}$ and to put $S=\frac{1}{2}$. This gives

$$(J=K-\tfrac{1}{2}):F=F_0{}^{el}+F^v+B_v\left[K(K+1)-\Lambda^2+\overline{G^2}\right]$$
$$-A\Lambda^2/2K-\gamma(K+1)/2+\phi_i(K,-\tfrac{1}{2})+\cdots \tag{43}$$

$$(J=K+\tfrac{1}{2}):F=F_0{}^{el}+F^v+B_v\left[K(K+1)-\Lambda^2+\overline{G^2}\right]$$
$$+A\Lambda^2/2(K+1)+\gamma K/2+\phi_i(K,+\tfrac{1}{2})+\cdots \tag{44}$$

Eqs. (43) and (44) now respectively agree with Eqs. (32, 34) and (32, 35).

For $A=0$, Eq. (37) takes the form which can be obtained by setting $A=0$ in Eqs. (41) or (42). Rather surprisingly, Eq. (37) takes the *same* form when $A/4B_v=+4$, as can be verified by direct substitution in Eq. (37). For this reason the relations of case b are closely approximated when $A/B_v\sim 4$ as well as when $A/B_v\sim 0$. The arrangement of the rotational energy levels is exactly the same for $A/B_v=+4$ as for $A/B_v=0$ except for the two lowest levels. When $A/B_v=0$ these form a close pair $[F_1(\Lambda+\frac{1}{2})$ and $F_2(\Lambda-\frac{1}{2})]$,

but when $A/B_v = +4$ the two lowest levels $[F_1(\Lambda - \frac{1}{2})$ and $F_1(\Lambda + \frac{1}{2})]$ are entirely separate, and behave like case a levels (cf. Eq. (38) in regard to the state having $J = \Lambda - \frac{1}{2}$). The difference between the two cases can best be seen by reference to Fig. 15, where the BeH levels correspond to $A/4B_v \sim 0$, while the arrangement of the MgH levels ($A/4B_v = +5.7$) is not very far from that characteristic of $A/B_v = +4$.

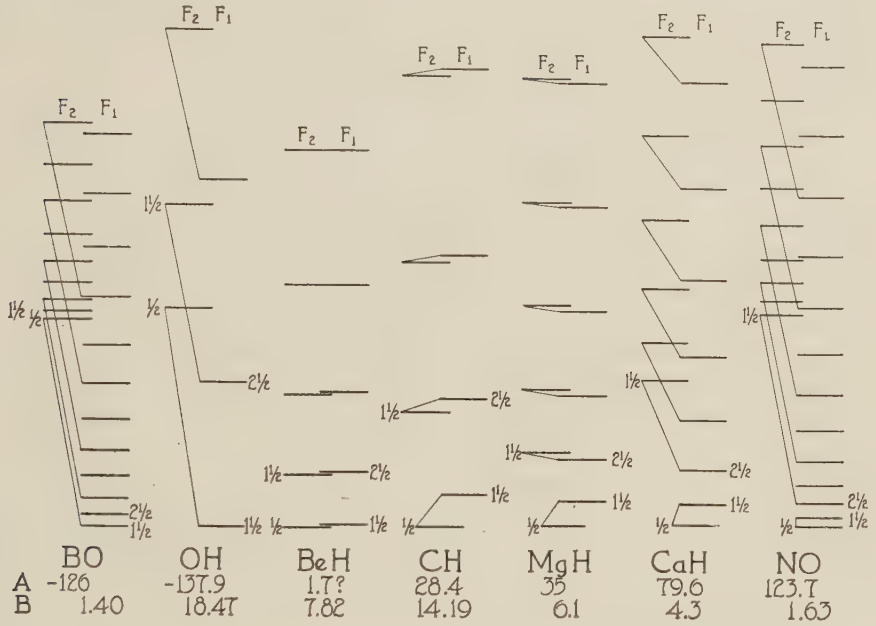


Fig. 15. Rotational energy levels, neglecting Λ-type doubling, for a series of $^2\Pi$ states with a variety of values of B_v and A/B_v (cf. R. S. Mulliken, Phys. Rev. **32**, 391 (1928)). The A values given have been recalculated (cf. R. S. Mulliken, Phys. Rev. **33**, 744 and 747 (1929)). The B values and rotational levels all correspond to $v=0$, except that in the case of NO they are for $v=4$. All the levels are on a uniform scale, with the lowest level placed at zero in each case. The numbers given opposite some of the levels are J values. Pairs of levels (F_1 and F_2) which would in case b correspond to the same value of K are indicated by slanting connecting lines. In such pairs, $F_2 > F_1$ when $A/B_v < 0$ or $> +4$, but $F_1 > F_2$ for A/B_v values between 0 and +4 (cf. CH in Fig. 15); corresponding F_1 and F_2 levels fall together when A/B_v is either ~ 0 or $\sim +4$ (cf. BeH for $A/B_v \sim 0$ and MgH for A/B_v near +4).—The level $F(\frac{1}{2})$ is classified as F_2 for $A/B_v < +2$, and as F_1 for $A/B_v > +2$. [Cf. text below, p. 112, for definition of F_1 and F_2 levels.]

Midway between the cases $A/B_v = 0$ and $+4$ is the characteristic case $A/B_v = +2$. Here Eq. (37) reduces to

$$F = F_0^{el} + F^v + B_v\left\{(J+\tfrac{1}{2})^2 - \Lambda^2 \pm \left[(J+\tfrac{1}{2})^2 - \Lambda^2\right]^{1/2} + \overline{G^2}\right\} + \phi_i(J) + \cdots. \quad (45)$$

This case is exemplified in the $^2\Pi$ normal state of CH (cf. Fig. 15). Eq. (45) is, accidentally, of practically the same form as, although of different significance than, the Kramers and Pauli equation which formerly was much used in representing the rotational energy functions of molecules.

Multiplet states: correlation of case a with case b energy levels. In case *a* doublet states we have two distinct sets of rotational levels, corresponding to different orientations of the spin; in the one set, $\Sigma=+\frac{1}{2}$ and $\Omega=\Lambda+\frac{1}{2}$, in the other $\Sigma=-\frac{1}{2}$ and $\Omega=\Lambda-\frac{1}{2}$. In case *b*, however, the two orientations of the spin, which give rise to $J=K\pm\frac{1}{2}$, correspond only to narrow doublets. The way in which the orientation of the spin would change in passing from normal case *a* ($A\gg0$) through case *b* ($A\sim0$) to inverted case *a*($A\ll0$) can be followed by a comparison of Eq. (37) with Eqs. (39)–(40) and (41)–(44). For $A\gg0$, the upper (+) sign of Eq. (37) gives Eq. (39), with $\Sigma=+\frac{1}{2}$, corresponding to $\Omega=\Lambda+\frac{1}{2}$; for $A\sim0$, it gives Eqs. (41)–(42), and (43), with $J=K-\frac{1}{2}$, and for $A\ll0$, it gives Eq. (40), with $\Sigma=-\frac{1}{2}$, corresponding to $\Omega=\Lambda-\frac{1}{2}$. Similarly, the lower (−) sign of Eq. (37) corresponds to $\Omega=\Lambda-\frac{1}{2}$ for $A\gg0$ (Eq. 39), to $J=K+\frac{1}{2}$ (Eqs. 41, 42, 44) for $A\sim0$, and to $\Omega=\Lambda+\frac{1}{2}$ for $A\ll0$ (Eq. 40). Thus if, for any specified J value, the spin is roughly speaking parallel (antiparallel) to Λ in inverted case *a*, it becomes parallel (antiparallel) to K in an imagined gradual change to case *b*, but then becomes antiparallel (parallel) to Λ when one passes on to normal case *a*.[48] [In saying that the spin is roughly speaking parallel (or antiparallel) to Λ, we mean merely that $\Sigma=+S=+\frac{1}{2}$ (or $=-S=-\frac{1}{2}$); actually $\boldsymbol{S}^*$ precesses around Λ with an angle given by $\cos\theta=\Sigma/S^*=(\pm\frac{1}{2})/(\sqrt{3}/2)$. Likewise in saying that the spin is parallel, or antiparallel, to K we mean merely that $J=K+S$, or $J=K-S$.] The single rotational state $J=\Lambda-\frac{1}{2}$ behaves in an exceptional manner, however (cf. Eq. (38)): the spin is antiparallel to Λ in both normal and inverted case *a*, and to K in case *b*. All these relations are indicated in Fig. 16 (cf. also Fig. 15). Fig. 16 also shows the relations between the Λ-type doublets of cases *a* and *b*.

In triplet and other states the correlation of energy levels between normal case *a*, case *b*, and inverted case *a* is analogous to that which holds for doublet states. In triplet states, for example, ($S=1$, $\Sigma=0,\pm1$), states with $\Omega=\Lambda\pm1$ in inverted case *a* correspond to case *b* states with $J=K\pm1$, and these in turn to states with $\Omega=\Lambda\mp1$ in normal case *a*; while $\Omega=\Lambda$ in inverted case *a* corresponds to $J=K$ in case *b* and to $\Omega=\Lambda$ in normal case *a*.

Classification of rotational levels in multiplet states. The above correlations serve as a convenient basis of classification. The writer has earlier proposed to designate as F_1 levels those levels which correspond to $J=K+S$ in case *b*, as F_2 levels those corresponding to $J=K+S-1$, and so on.[48] Thus for doublets one has F_1 and F_2 levels ($J=K\pm\frac{1}{2}$), for triplets F_1, F_2, and F_3 levels ($J=K+1$, K, $K-1$, respectively), and so on. The classification suggested is useful not only for case *b* but also for intermediate cases between *a* and *b* and even for case *a*. Taking $^2\Pi$ states as an example, the $^2\Pi_{\frac{1}{2}}$ and $^2\Pi_{1\frac{1}{2}}$ states, respectively, of normal case *a*, and the $^2\Pi_{1\frac{1}{2}}$ and $^2\Pi_{\frac{1}{2}}$ states, respectively, of inverted case *a* would be classified, the former as F_1, the latter as F_2 states[49]

[48] Cf. R. S. Mulliken, Phys. Rev. **30**, 793–6 and Fig. 1 (1927) for a further discussion.

[49] In doublet states, the level having the special J value $J=\Lambda-\frac{1}{2}$ (e.g. $J=\frac{1}{2}$ in $^2\Pi$ states) is classified as an F_2 level for $A<2B_v$ and as an F_1 level for $A>2B_v$, corresponding to the fact of its obvious association with the $^2\Pi_{1/2}$ levels in both normal and inverted case *a* (cf. Eq. (38) and Fig. 15).

(cf. Figs. 15, 16). Similarly, normal case a $^3\Pi_0$, $^3\Pi_1$, $^3\Pi_2$ states, and inverted case a $^3\Pi_2$, $^3\Pi_1$, $^3\Pi_0$ states, would be respectively classified as F_1, F_2, F_3. Fig. 15 shows the application of the above classification to a series of typical $^2\Pi$ molecular states.

In addition to the classification F_1, F_2, $\cdots$, which indicates the orientation of the spin, the two sub-levels a and b which occur (Λ-type doubling) for each such orientation and for each J value, may be designated by additional subscripts, thus, F_{1a}, F_{1b}, F_{2a}, F_{2b}, $\cdots$.

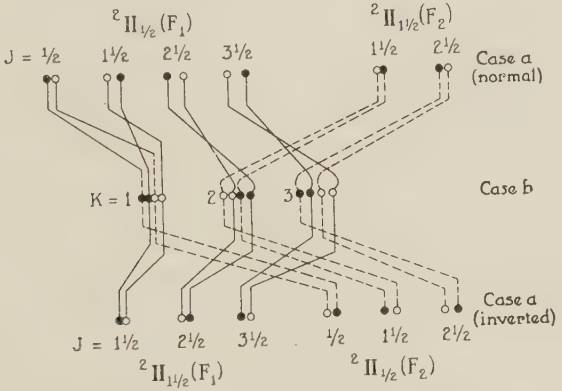


Fig. 16. Correlation of rotational energy levels between Hund's cases a and b (cf. Van Vleck, p. 496 of reference 43). F_1 levels are connected by means of full lines, F_2 levels by means of dotted lines. Λ-type doublets are shown by the use of full and empty circles (● and ○). In both normal and inverted case a, the Λ-type doublets are, according to theory and experiment (reference 43), much wider for $^2\Pi_{\frac{1}{2}}$ than for $^2\Pi_{1\frac{1}{2}}$ states; this is shown in a *qualitatively* correct way in the figure. The crossings of levels belonging to Λ-type doublets, between cases a and b, are as given by Van Vleck.[43] In case b, the Λ-type doublets are of the same width for F_1 and F_2 states; both these and the spin doublets are shown, with exaggerated spacings, in the figure; the spin doublets are drawn in qualitative agreement with the usual relation that the energy is higher for the F_1 than for the F_2 component; the relative spacings of the spin and Λ-type doublets as given for case b in the figure are, however, not significant. The various indicated crossings of corresponding F_1 and F_2 levels (same K, different J) are in agreement with theory and experiment (cf. Fig. 15). [Cf. text, p. 112, for definition of F_1 and F_2 levels.]

Doublet states: comparison of theoretical equations with experimental data for $^2\Pi$ states. For a comparison between Hill and Van Vleck's doublet formula (Eq. (37)) and experiment, extensive data are available only for $^2\Pi$ states. Fig. 15 shows the observed energy levels for a number of such states, corresponding to a wide range of A values. In most of the examples in this figure, A is large enough so that the approximations given by Eqs. (39) and (40) hold, at least for small J values. If we let C_Σ represent a suitable (very small) constant whose value depends on Σ, and if we set

$$B_{v,\Sigma}^* = B_v(1 \pm B_v/A\Lambda), \tag{46}$$

with $\pm B_v/A\Lambda$ according as $\Sigma = \pm \frac{1}{2}$, and if we substitute $\Omega = \Lambda + \Sigma$, we get from both Eqs. (39) and (40)

$$F=F_0{}^{el}+F^v+A\Lambda\Sigma+C_\Sigma+B_{v,\Sigma}{}^*[J(J+1)-\Omega^2+(\tfrac{1}{2})+\overline{G^2}]+\phi_i(\Sigma,J)+\cdots. \quad (47)$$

As B_v/A becomes smaller, the two values of $B_{v,\Sigma}{}^*$ both approach B_v, while C_Σ vanishes, and Eq. (47) goes over into Eq. (31). Eqs. (46) and (47) give a satisfactory explanation of the experimental fact[50] (cf. Fig. 15) that in $^2\Pi$ states, if $|A|/B_v$ is fairly large, the rotational levels obey a relation of the *form* of Eq. (31) but with a larger coefficient $B_v{}^*$ for the F_2 than for the F_1 levels. In agreement with the theory, the difference between the $B_v{}^*$ values for the F_2 and F_1 sets of levels is large for the smaller values of $|A|/B_v$ and conversely. Also in agreement with the theory is the fact that $B_v{}^*$ is always greater for the F_2 levels, whether $A<0$, even though these correspond to $^2\Pi_{\frac{3}{2}}$ levels if $A<0$ but to $^2\Pi_{1\frac{1}{2}}$ levels if $A<0$.

According to Hill and Van Vleck's equation (Eq. (37)) the region of A/B_v values from 0 to $+4$ should show some peculiarities. In this region $F_1(K)>F_2(K)$, whereas elsewhere $F_2(K)>F_1(K)$. The relation $F_1(K)>F_2(K)$ is actually found in the $^2\Pi$ state of CH (cf. Fig. 15 and refs. 48 and 50), for which $A/B_v=+2$ (cf. discussion following Eq. (45)).

Other coupling cases. Besides Hund's cases *a*, *b* and *d*, there are other conceivable modes of coupling of the electronic and nuclear angular momentum vectors. In *Hund's case c*[24] we have $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ coupled to give a resultant $J^a{}^*$, which then gives a projection Ω on the electric axis, exactly as in the case of an atom in a weak electric field (cf. Fig. 10). We use J^a instead of J here for the quantum number corresponding to the resultant of $\boldsymbol{L}^*$ and $\boldsymbol{S}$, reserving J as usual for the quantum number corresponding to the resultant angular momentum of the molecule. For the rotating molecule, $\boldsymbol{\Omega}$ and $\boldsymbol{O}$ form a resultant $\boldsymbol{J}^*$, just as in case *a* (cf. p. 105), and the energy is given by an expression of the form

$$F=\{F_0{}^{el}+(\tfrac{1}{2})A[J^a(J^a+1)-L(L+1)-S(S+1)]+F(\Omega)\}+F^v$$
$$+B_v[J(J+1)-\Omega^2+\overline{H^2}]+\phi_i(J)+D_vJ^2(J+1)^2+\cdots. \quad (48)$$

Here H represents the component of electronic angular momentum (including both orbital momentum and spin) perpendicular to the electric axis; $\overline{H^2}$ replaces $\overline{G^2}+S_{perp}{}^2$ of Eq. (31). Hund's case *c* is not yet known in practise. This is because the effect of the electric axis in the case of known molecular states is strong enough to break down the coupling of $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ practically completely. We may, however, expect to find examples of Hund's case *c* among the less stable states of molecules containing heavy atoms (for heavy atoms, the coupling of $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ is relatively large).

For very rapid rotations, case *c* should pass continuously into another case, which we may call case *e*, in a manner analogous to the passage of case *b* into case *d*.[36a] In case *e* we assume that $\boldsymbol{L}^*$ and $\boldsymbol{S}^*$ give $\boldsymbol{J}^a{}^*$ as in case *c*, but that the rotation of the nuclei is so rapid, or the coupling of $\boldsymbol{J}^a$ to the electric axis so small, that Ω no longer exists. Instead we have the nuclei rotating with a quantum number R as in case *d*, then $\boldsymbol{J}^a{}^*$ and $\boldsymbol{R}^*$ forming a resultant $\boldsymbol{J}^*$ around which they precess. Case *e* has not yet been found in practise.

[50] Cf. R. S. Mulliken, Phys. Rev. **32**, 389–392 (1928).

Summary of Hund's cases. The various types of coupling which have been discussed above are summarized in Table IV. Under "Projection and composition of vectors" are given first, under "Elements," the symbols corresponding to those quantum vectors or projections of quantum vectors which serve as elements in the building up of the total angular momentum

TABLE IV. *Summary of Hund's Coupling Cases*

Case	Coupling of L^* to electric axis	Coupling of S^* to L^*	Effect of rotation (on S^* or L^* or both)	Projection and composition of vectors		Limitations
				Elements	Composition	
a	large	moderate	small	$L^*, \Lambda; S^*, \Sigma,$ with $\Omega = \lvert \Lambda+\Sigma \rvert$	$\Omega+O(=L^*+S^*+N)=J^*$	$\Lambda>0, S>0$
b	large	small	moderate	$L^*, \Lambda; S^*$	$\Lambda+O(=L^*+N)=K^*;$ $K^*+S^*=J^*$	
d	moderate	small	large	$L^*; S^*; R^*$	$R^*+L^*=K^*, K^*+S^*=J^*$	
c	moderate	large	small	$L^*+S^*=J^{a*};$ J^{a*}, Ω	$\Omega+O(=J^{a*}+N)=J^*$	$L>0, S>0$
e	small	moderate	large	$L+S=J^{a*}; R^*$	$J^{a*}+R^*=J^*$	$L>0, S>0$
b', d'	Same as *b*, *d*, but $S=0$, or coupling of S^* to K^* negligibly small.					

for the case under consideration; when the projection of a vector on the electric axis is important, the vector and its projection are both given, as in the example L^*, Λ. Under "Composition", equations are given, which are to be understood as vector equations, showing how the electronic and nuclear angular momentum vectors are compounded.

As a result of a careful study, made since this review was written, the writer recommends the following changes of notation as compared with the present text; E_e, E_v, and E_r instead of E^{el}, E^v, and E^r; T, T^e, G, and F instead of F, F^{el}, F^v, and F^r; $\Delta G(v)$, defined as $G(v+\frac{1}{2})-G(v-\frac{1}{2})$, instead of $\Delta F^v(v_1, v_2)$.

The substance of these articles, revised and with additional material, will be subsequently published in book form. The writer will be very grateful to anyone who will call his attention to errors and obscure passages in the present treatment.

The reader will find in Ruark and Urey's new book "Atoms, Molecules, and Quanta" (McGraw-Hill, 1930) an extremely valuable treatment of the subject of molecular spectra.

X-RAY DIFFRACTION IN LIQUIDS

By G. W. Stewart
University of Iowa

WHAT is the cause of halos formed by x-rays passing through supposedly amorphous substances such as liquids?[1] Surely the liquid does not contain minute crystals, and if not, what causes the effects obtained? These queries were not satisfactorily answered three years ago. Today it is agreed, at least by the most active investigators and by all who are publishing opinions, that the diffraction effects with liquids do indicate spatial orderly arrangements of molecules. This view may seem at first a surprising admission, but, upon careful consideration, it becomes most reasonable. The purpose of this paper is to elaborate this view and to give the outline of the experimental results which substantiate it. The status of the study of x-ray diffraction in liquids is briefly this: (1) the phenomenon is understood qualitatively in the same sense that we understand crystal diffraction; (2) theoretical analysis has not given great assistance because of the complexity in liquids and the corresponding necessary crude approximations; (3) the value of x-rays in the study of the nature of liquids, of solutions and of molecular structure has been demonstrated.

X-ray diffraction has proved to be a powerful means for the study of crystal structure. The diffraction law of Bragg, $\lambda = 2d \sin \theta/2$, has won full confidence by the consistent results obtained through its use. But its correctness has been actually proved. For the values of the wave-length, λ, as determined from measurements with crystals, have been verified by experiments with ruled line gratings. It is the periodicity of the crystal planes, separated by the distance d that gives the diffraction maxima at angles $\theta/2$, where θ is the deviation of the x-rays, and this is the view of both the electromagnetic and quantum theories. Throughout this paper, however, the language will be in terms of the electromagnetic theory. In the crystal the periodicity is not that of diffraction centers located in actual planes, for the x-rays are scattered by the electrons which are distributed throughout the atoms and consequently cannot all lie in planes. If the density of distribution varies periodically with depth in a crystal, then theory shows that there will be regular reflection of x-rays in accord with Bragg's formula if d is the distance of separation of recurring densities. If the distribution concentrates the

[1] This article does not refer to liquid crystals but to optically isotropic liquids. The connection of our discussion with the condition called "liquid crystals" is quite definite and is chiefly one of magnitude of the individual molecular groups. But the discussion of this point is beyond the scope of this article. The discussion is confined to what are ordinarily regarded as "amorphous" liquids. It is also the policy of this article to omit the details of the argument favoring its interpretation of the present status, emphasizing rather the nature and diversity of results.

electrons in planes, then the intensity of the diffraction at the angle $\theta/2$ is large. As, however, the distribution of electrons approaches a uniform one between these planes, the regular reflection approaches a zero value. Any distribution that is roughly periodic in space would give an intensity maximum for the λ and θ of Bragg's formula.

From these considerations it is obvious that if there is a sufficiently marked space periodicity in the structure of liquids, there would be found an x-ray diffraction effect. Debye and Scherrer[2] were the first observers of the diffraction halo in liquids. The term "halo" is used because the experiments were performed with a circular beam of x-rays and the diffraction ring thus formed is much broader and more diffuse than would be formed with a powdered crystal. It is agreed by all that this halo demonstrates a periodicity in the liquid, but its nature has been the cause of discussion.

Is the effective periodicity located within the atom, (electron distribution in the atom), within the molecule (atomic distribution), or intermolecular (molecular distribution)? The answer is to be found through experiment. Even then the three possibilities can be reduced to two or one only in a qualitative sense, for all the effects may exist although not necessarily detectable. Doubtless the halo found in monatomic substances[3] should be accounted for by the first cause. When complex molecules are used, the third cause is the predominating one, since the chief diffraction maxima are accounted for by a periodicity in the distribution of molecules. (Of course it is understood that the ultimate diffraction centers are electrons, and that, with the correct regard for phase differences those of one atom may be considered as grouped at atomic "centers" and that, again by a similar process, the electrons of all the atoms in a molecule may be regarded as grouped at a molecular "center." Thus three interference effects, between electrons in the atom, between atoms and between molecules, are based upon the scattering of x-rays by electrons.)

All three types of distribution of diffraction centers are of interest, for we desire to secure information leading to any evidence as to arrangements of electrons in atoms, of atoms in the molecule, and molecules in any group of molecules. But with liquids in general, experiment shows that the interest must rest chiefly in the last named, for the results thus far obtained are not capable of giving a considerable amount of finer detail.

What can cause a molecular grouping that is perfect enough to lead to diffraction halos? The reply which might come to mind at once is that the "shape" of the molecules, if sufficiently asymmetrical, might cause such an arrangement. Thus if the molecules were cigar shaped, the probability of parallel grouping of the molecules would be greater than that of any other. But shape has no meaning other than the space apparently occupied by the molecules and the molecular grouping would depend upon the nature of the

[2] Debye and Scherrer, Nachr. Göttingen, 1916.

[3] Keesom and de Smedt, Versl. Akad. Amsterdam **31**, 87 (1922) and **32**, 72 (1923), used argon.

molecular forces. In the crystal, the forces are electrostatic and electrodynamic, and the atoms (or molecules) are in stable configuration in mean positions representing minimum potential energy. If then in a liquid, molecules, oblong in shape, have a high probability of parallelism, it is because of the nature of the molecular forces. And it is not at all surprising if oblong molecules are thus frequently found parallel. The kinetic energy possessed by the molecules because of specific heat will not necessarily disturb the parallelism to a serious extent. For this kinetic energy is represented by the acoustic waves which travel throughout the liquid, permitting the relative velocities of two adjacent molecules to be small in comparison with the actual velocity of either.

If the x-rays give evidence of periodic molecular grouping, it must not be supposed that these groups are large or that the molecules in any one well defined group remain permanently members of that group. At any one instant these small orderly molecular groups might exist at numerous points in the liquid, the regions between them being not so orderly. The percentage of the volume occupied by fairly orderly groups might be large in some liquids and not so large in others. The evidence today strongly favors this orderly arrangement in groups which has been called[4] the "cybotactic" condition. This view will be discussed at greater length in the latter part of the article.

With the foregoing introductory remarks concerning the nature of the phenomena involved, an account will now be given of some of the results obtained. Reference will not be made to all the literature, for the endeavor of this article will be to give to the reader the present status of experimental results without burdening him with all the material now in print.[5] For a partial list of the literature up to 1928, the reader is advised to consult Drucker's article.[6]

The diameters of chain molecules. In the diffraction rings of chain molecules, such as n-alcohols, n-fatty acids, and n-paraffins there is always a major intensity maximum which is located at approximately the same angular deviation. It indicates the separation of planes d, by approximately 4.6 Å.u. The interpretation made by the investigators[7] is that d is the effective "diameter" of these chain molecules. The square of this value, 4.6 Å.u., is the area of cross section occupied by the chain. This agrees with the well-known surface film experiments of Adam. It should be emphasized at this point, however, that the measurements of observers who have used different x-ray beams for examination must of necessity be expected to differ slightly in computed results. For, unless the radiation is strictly monochromatic.

[4] Stewart and Morrow, Phys. Rev. **30**, p. 232 (1927).

[5] There have been published approximately sixty papers containing data on x-ray diffraction in about two hundred fifty different liquids.

[6] Drucker, Phys. Zeits. **29**, p. 273 (1928).

[7] Sogani, Ind. Jl. Physics **1**, p. 357 (1927); Katz, Zeits. f. Physik **45**, p. 97 (1927); Vaiyanathan, Ind. Jl. Physics **3**, p. 391 (1928); Stewart and Morrow, Phys. Rev. **30**, p. 232 (1927); Morrow, Phys. Rev. **31**, p. 10 (1928); Stewart, Phys. Rev. **32**, p. 153 (1928).

the diffraction angle of maximum intensity may not correspond exactly to the wave-length used in computations. For this reason it will be the policy of this paper to ignore these minor differences in distances measured by the workers in this field. Just now the experimental work is largely of a scouting nature. At a later time, more accurate measurements will be made where needed. Then details of disagreements in results will be in point.

Effect of branches upon the diameter of chain molecules. Isomers of the n-alcohols and of n-paraffins have been examined[8] and the branches increase the effective diameters of the chains. The details contain certain information in regard to the structure of molecules, but not sufficiently definite for this resumé.

Existence of double molecules in polar compounds. In polar compounds, especially in n-alcohols, n-fatty acids, and benzene and cyclohexane derivatives, there is frequently a second intensity maximum in the x-ray diffraction, which, using Bragg's law, gives a d greater than the length of one molecule, in some cases equal to twice the molecular length and in others less than twice this length. There seems to be no objection to the conclusion that this measurement indicates the grouping of two polar molecules by the attraction of the two polar groups. With the chain molecules this means an arrangement of the two attracting polar molecules end to end in a line and parallel to similar neighbors. With the chain molecules and an OH group, if the latter is attached in a branch, the end to end arrangement is replaced by a parallel side by side arrangement of molecules, provided the attachment is elsewhere than next to the end carbon atom.

Nature of the carbon chains. Experiments[9] with the n-paraffins, the assumption of the parallel arrangement of the molecules and the best values of the densities, lead to the conclusion that the carbon atom occupies a distance of 1.24 Å. u. along the chain. This is in agreement with results with alcohols when the length of single molecules are measured. This distance along the chain compared with separation in the diamond, 1.54 Å. u. and the diameter of the chain 4.6 Å. u. leads one to favor a nonlinear arrangement of the carbon atoms in the chains. There is found other corroborative evidence such as a decrease in molecular diameter with decrease in the number of carbon atoms in case there are only from one to five in the chain.

Simultaneous measurement of more than one diameter of a chain. This has been accomplished in two cases, 2-methyl hexane[10] and di-n-propyl carbinol.[11] The measurements were 5.25 Å.u. and 4.84 Å.u. for the former and 4.85 and 4.5 Å.u. for the latter. In the latter there was also a third intensity maximum from which the length of a single molecule was obtained. Such evidence, and there is much of it, shows that the diffraction intensities in-

[8] Chiefly by Stewart and Skinner, Phys. Rev. **31**, p. 1 (1928); Stewart, Phys. Rev. **32**, p. 153 (1928).

[9] Stewart, Phys. Rev. **31**, p. 174 (1928).

[10] Stewart, Phys. Rev. **31**, p. 174 (1928).

[11] Stewart, Phys. Rev. **31**, p. 1 (1928).

dicate actual molecular arrangement. One could not get three such mean or average values for one liquid.

Benzene, cyclohexane and their derivates. The benzene and cyclohexane "rings" give unusually sharp diffraction intensity variations. There is at present a difference of interpretation of the measurements on these compounds and their derivatives. Krishnamurti[12] interprets his work as indicating the thickness of the ring for benzene to be about 3 Å.u. The author,[13] with a larger range of material, favors the view that benzene and cyclohexane have a thickness of 4.7 Å.u. and 5.1 Å.u. respectively, and corresponding areas perpendicular to the thickness of 31.4 Å.u. and 35.2 Å.u. The area of the former thickness agrees with crystalline determinations, but its thickness is much greater, 4.7 Å.u. as contrasted with 3.0 Å.u. in the crystal.[14]

The *o*—, *m*—, and *p*— drivatives of benzene and cyclohexane have decreasing thicknesses, the para compound beng the least. This has been shown only for five series of compounds. (See reference (12)). The data are too meager for generalization.

Solutions. Krishnamurti[15] has investigated the cases of ammonium nitrate and acetamide in water, different kinds of sugar solutions, aqueous solutions of phenol, trimethyl carbinol, piperidine, glycerine, ethyl alcohol, lactic acid, pyridine and butyric acids. Hertlein[16] used aqueous solutions of butyl and amyl alcohol. Prins[17] studied the effect of heavy atoms dissolved in liquid. Banerje[18] investigated x-ray diffraction in liquid alloys of sodium and potassium. The results of all these observations lead us to understand that the investigations by means of x-ray diffraction emphasize differences in the nature of solutions. With one class of solutions, the diffraction maximum intensities do not correspond to those of the component liquids taken separately, but the solution seems to be a *new* space arrangement. Such liquids really hold each other in an intimate manner. Another class of solutions exist in which each of the two components shows its own intensity maximum, indicating an independent space grouping, each with its own kind. Then there are solutions which can be put in neither class. It is increasingly evident that the x-ray experiments do shed light upon the nature of solutions and will become an interesting method of their study. It has already shown that with some solutions the nature of the binding forces between the two components is similar to those existing in the crystalline structure of solid solutions.

Effect of temperature. So far as experiments have proceeded, the effect of increasing the temperature is very similar to that in the case of the crystal.

[12] Krishnamurti, Ind. Phys. **3,** p. 275 (1929).

[13] Stewart, Phys. Rev. **33,** 889 (1929).

[14] There is no reason to doubt the measurements with either crystals or liquids. The effective diameter may be greater in the liquid and the differences of method may also cause a difference.

[15] Krishnamurti, Ind. Jl., Phys. **2,** p. 501 (1928) and **3,** p. 209 (1928) and **3,** p. 331 (1928).

[16] Hertlein, Zeits. f. Physik **54,** p. 341 (1929).

[17] Prins, Nature, p. 84, Jan. **19,** 1929.

[18] Banerje, Ind. Jl. of Phys. **3,** p. 399 (1929).

There is a displacement of the intensity maximum, a diminution of the magnitude of this maximum and a broadening of the maximum peak in an intensity diffraction curve. This similarity to effects in a crystal adds weight to the conception of the cybotactic condition.

Theory of x-ray diffraction in liquids. There are two ways of approach to an adequate theory of x-ray diffraction in liquids. One is by assuming ideal systems, such as molecules each of which occupies a spherical space. The other approach is by the consideration of molecules that have asymetrical forces and occupy spaces not spherical. The former leads more easily to a mathematical expression. The latter is more descriptive of the actual case. Three theoretical contributions of the former type will be mentioned. Ehrenfest[19] considered the scattering produced by pairs of electrons separated by a distance a, and assumed the pairs to be independent. Thus his theory was applicable to gases only. Debye[20] considered the scattering in gases both with single scattering particles and pairs. He considered the results with the molecules vanishingly small and with them occupying a considerable portion of the total volume. Raman and Ramanathan[21] discussed liquids and the effect of thermal fluctuations of density and applied probability considerations to the liquid as a continuum. They assumed symmetrical molecules. Their theory is the one most applicable to a liquid and is in agreement with observations in several respects.

But the most direct approach to an understanding is the descriptive rather than the analytical theory and is presented in the introductory remarks. The foregoing results of computation from x-ray diffraction, varied in nature, seem clearly to give confidence in the assumption of the cybotactic condition in liquids in general and in the validity of the use of Bragg's formula.[22] There is undoubtedly a spatial arrangement of molecules so general that throughout the liquid at any instant can be found frequent, well ordered groups. Quantitative measurements, not here described, indicate that most of the liquid is effective in producing marked diffraction effects. In fact, the liquid will produce coherent diffraction practically as intense as the powdered crystal of the same material. This description is what one would always have naturally supposed, were it not for the kinetic heat energy. The equivalence of a continuum and acoustic waves is, however, a satisfactory assumption. We are to regard the cybotactic condition as a common one, and the forces involved more asymmetrical then even the shape of the space occupied by

[19] Ehrenfest, Proc. Akad. Wet. Amsterdam **27**, p. 1184 (1915).

[20] Debye, Jl. of Mathematics and Physics **4**, p. 133 (1925) and in German in Phys. Zeits. **28**, p. 135 (1927).

[21] Raman and Ramanathan, Proc. Ind. Assoc. for Culti. Sci. **8**, p. 127 (1923).

[22] In using Bragg's Law with the foregoing determinations of effective diameters we are making two assumptions which must be borne in mind, though, for present purposes, they are nearly enough correct. We choose to regard the d, computed for the maximum in the diffraction intensity-angle curve, as the separation of diffraction centers, when the curve is really a statistical result. Also we have assumed a very simple picture of molecular arrangement in interpreting a given d as the effective molecular diameter.

a molecule would lead one to believe. The theory from a descriptive view point, is then satisfactorily established.[23] On the analytical side, however, there is much room for improvement. In the meantime experiment in the field will add to our knowledge of molecular structure and the nature of solutions.

[23] Krishnamurti, (in a very recent paper published in reprint form May 1, 1929, and to appear in Ind. Jl. of Physics, **3** 1929) goes a little further and indicates that the spatial arrangement is similar to that in crystals, in fact that "the planes containing the maximum number of molecules per unit area . . . . occur in the liquid approximately as they do in a crystal."

Volume 2 *April, 1930* *Number 2*

REVIEWS
OF
MODERN PHYSICS

ELECTRICAL DISCHARGES IN GASES

PART I. SURVEY OF FUNDAMENTAL PROCESSES

By
KARL T. COMPTON, *Princeton University*
and
IRVING LANGMUIR, *Research Laboratory of the General Electric Company*

TABLE OF CONTENTS

A. PRODUCTION OF ELECTRONS AND IONS IN GASES

(1) **Ionization by electron impact** 124
Critical potentials, probability of ionization, products of ionization.
(2) **Photo-ionization** 129
Probability of photo-ionization as function of frequency.
(3) **Cumulative ionization** 131
Successive impact, successive absorption, photo-impact, resonance radiation, metastable states, average life times of excited and metastable states.
(4) **Ionization by positive ion impact** 134
(5) **Ionization by collisions of the second kind** 135
(6) **Thermal ionization** 135
Saha's equation and the more general form suitable to a laboratory device.

B. LIBERATION OF ELECTRONS AND IONS FROM ELECTRODES

(1) **Thermionic emission of electrons** 136
Relation with temperature and nature of metal, work function, assumptions underlying theories, activated surfaces, emission velocities, heat of evaporation of electrons.
(2) **Thermionic emission of positive ions** 140
(3) **Photo-electric emission of electrons** 141
Threshold frequency, velocities of emission, Einstein's equation, cumulative emission, total emission, emission under influence of black body radiation, thin films, selective photo-electric effect.
(4) **Relations of contact difference of potential** 144
Thermionic, photo-electric work functions and contact potentials, distinction between homogeneous and patchy surfaces.
(5) **Electron emission in an accelerating electric field** 146
Schottky effect, surface forces, meaning of saturation emission, peculiarities of non-homogeneous and activated surfaces, theories of "grid" action of patches and of adions.
(6) **Electron emission due to high fields** 160
Field currents, effect of rough surfaces, dependence on temperature and field, local heating.

(7) **Electron emission due to electron bombardment and electron reflection** 171
Secondary or "delta ray" emission, variation with primary velocity, velocities of emission, critical potentials, diffraction of electrons as waves.
(8) **Electron emission due to impact of metastable atoms** 175
(9) **Electron emission due to positive ion bombardment** 177
(10) **Emission due to chemical action** 181

C. Disappearance of Electrons and Ions

(1) **Disappearance at the electrodes and walls** 182
Heating effects, accommodation coefficient, momentum transfer, sputtering.
(2) **Disappearance by recombination** 191
Recombination on surfaces, recombination of ions, electron attachment, Langevin's theory, Thomson's theory, recombination of free electrons and ions, recombination spectra, variation of probability of recombination with velocity, theories and relation with photo-ionization, experiments of Davis and Barnes.

D. Motions of Electrons and Ions

(1) **Classical kinetic theory** 204
(a) Maxwell's and Boltzmann's distribution laws; (b) mean free paths and influence of molecular attraction; (c) collisions and single scattering in reference to energy loss and directions of scattering, cases of elastic spheres and of inverse power laws; (d) diffusion; (e) relation of diffusion to mobility, mobility, terminal speeds.
(2) **Comparison of classical kinetic theory with experiments** 221
(a) Velocity and concentration distributions; (b) mean free paths and Ramsauer effect; (c) nature of collisions and scattering of electrons at collisions with molecules; (d) electron and ion diffusion; (e) electron and ion mobilities.
(3) **Quantum theories of collisions and scattering** 237
Small deflections, angular distribution of scattering, Ramsauer effect.
(4) **Space charge oscillations** 239
Plasma electron oscillations, plasma ion oscillations.

As A result of continuous intensive investigation of the properties and phenomena of ionized gases during the past thirty years, it is likely that most of the fundamental elementary processes occurring in gas discharges are now recognized. The great problem at present is to analyze the complex phenomena so that they can be quantitatively described in terms of these fundamental processes in every region of each one of the widely differing types of gas discharges. These fundamental processes are classified and described briefly in this article. Some of them will be discussed more in Part II, which will deal primarily with a study of gas discharges themselves. In the present discussion particular attention is given to those aspects of the subject which are still rather obscure, and some phenomena which are very important are not discussed in such detail because they are quite familiar and generally understood. In every case, however, an attempt is made to give such references to the literature of the subject as will enable an interested reader easily to make his own more intensive study.

A. Production of Electrons and Ions in Gases

(1) **Ionization by electron impact,** originally suggested by Townsend,[1] appears to be of first importance in nearly all gas discharges. In order to

[1] Townsend, Phil. Mag. **1**, 198 (1901); "Electricity in Gases," Chap. VIII (Oxford 1915).

ionize any normal gas molecule,* the impacting electron must possess kinetic energy in excess of a certain minimum amount which is characteristic of each particular type of gas molecule. This minimum energy is usually specified as the *Minimum Ionizing Potential*[2] of the gas, since it is usually measured in terms of the potential difference V_i through which an electron must move in order to acquire this minimum ionizing energy. There are also higher ionizing potentials corresponding to the removal of two electrons simultaneously, or to the removal of a second after the first had previously been removed, or to the removal of one of the more deep-seated electrons from the molecule. Since the minimum energy required to remove an electron from a molecule is the same, whether this energy be supplied by an impinging electron (impact ionization) or by absorption of radiation (photo-ionization), we can equate these energies giving

$$eV_i = h\nu_i,$$

where ν_i is the convergence frequency of the spectral absorption series. This relation, whose proof and applications were so actively pursued by physicists between five and ten years ago, is now so well understood that no further discussion of it need by given here.

The *probability of ionization of a gas molecule by an electron impact*[3] apparently increases from zero approximately linearly with increase in energy of the impacting electron from the minimum ionizing energy up to about twice this energy, and then more slowly.[4] The experimental determination of this probability involves measurement first of the number of new ions formed per unit path by an electron of given energy and second of the measurement of the total number of collisions made per unit path by the electron. As we shall see, this latter quantity is not easy to measure nor even to define accurately. Thus, while it is often convenient to know approximately the proba-

* The term "molecule" is here used in its original sense, including monatomic as well as polyatomic molecules.

[2] For methods of determining ionizing potentials and their values see Compton and Mohler, "Critical Potentials," National Research Council Bulletin No. 192; Franck and Jordan, "Anregung von Quantensprüngen durch Stösse," (Springer, 1926); Geiger and Scheel, "Handbuch der Physik," Vol. XXIII (1927); Bloch, "Ionization et Resonance des Gas et des Vapeurs," (Société Francaise de Physique, 1925); Mohler, "Critical Potentials of Atoms and Molecules," International Critical Tables Vol. VI, pp. 69–74 (1929).

[3] Compton and Van Voorhis, Phys. Rev. **27,** 724 (1926) give values for a number of gases up to 400 volts. Jones, Phys. Rev. **29,** 822 (1927) and Bleakney, *ibid.*, **35,** 139 (1930) give values for Hg vapor. Buckmann, Ann. d. Physik **87,** 509 (1918) gives values for air up to electron energies of 25,000 volts. Penning, Physica **6,** 290 (1926) shows that, if actual experimental values of the numbers of electron collisions are used in computing the probability, then the probability of ionization *at a collision* is found for a number of gases to be about the same function of the electron energy, if the unit of energy for each gas is taken to be its minimum ionizing energy.

[4] This is contrary to conclusions drawn by von Hippel, Ann. d. Physik **87,** 1035 (1928) whose results seem impossible to reconcile with those of other observers and whose method of measurement seems to have allowed some unsuspected complicating factor to play an overwhelming role. See also, however, Funk, Ann. d. Physik **4,** 149 (1930).

bility of ionization at an impact, it is simpler for purposes of the theory of gas discharges to deal directly with the more accurately determined quantity, the number of new electrons produced by ionization per unit path at a specified gas pressure by an electron of given energy. This may be called the *probability of ionization per unit path at unit pressure.*[3]

If the energy of the impacting electron exceeds the amount necessary to disengage two or more electrons from the molecule, then it is necessary to consider the nature of these products of ionization, *i.e.* whether they are singly, doubly, triply, etc., ionized. This was first done by Smyth,[11] and very recently by Bleakney,[3] who has determined also the probability of production of each of the various types of Hg ions as a function of the energy of the impacting electrons. These results are shown in Fig. 1.

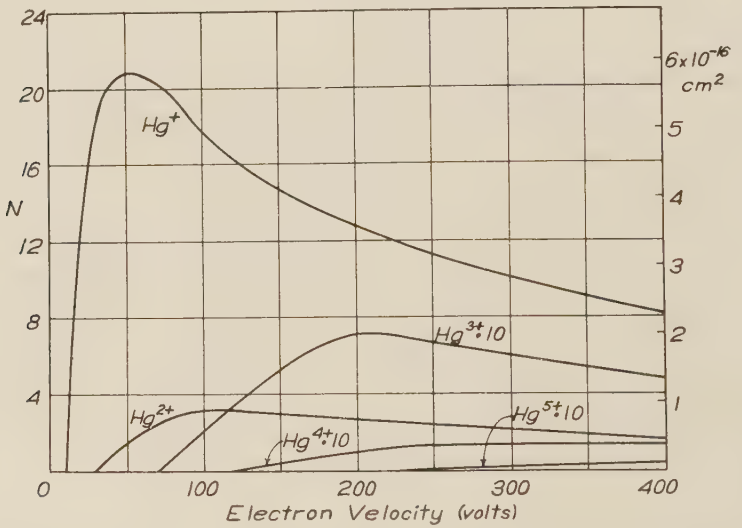


Fig. 1. Number of Hg^+, Hg^{2+}, Hg^{3+}, Hg^{4+} and Hg^{5+} ions per cm path per mm pressure at 0°C. To the right is the effective cross sectional area for ionization of a Hg atom.

The method by which the values of P in Table I were found is such that a doubly ionized molecule would be recorded as two single ions, a trebly charged molecule as three ions, etc. To the extent to which multiple ionization occurs by single electron impacts at the higher voltages, the values in Table I are in error *if considered in connection with any theory of ionization by impact.* For such considerations, separate measurements of single and multiple stages of ionization are desirable, as attempted by von Hippel, and achieved by Bleakney. But there is high probability that a multiply charged ion, at a collision with a neutral molecule, robs an electron from the molecule, so that very quickly the state is reached in which the multiply charged ions are replaced by singly charged ions of equivalent total charge. Furthermore, since the probability of forming original multiply charged ions appears always to be less than that of forming singly charged ions, it would seem

justifiable to take values of P such as those in Table I to be correct *if considered in connection with theories of gas discharges*, for here it is the total rate of production of ions which is of interest.

TABLE I. *Ionization of gases by electron impact.*

V_i = minimum ionizing potential.

$P(V)$ = probability that an electron with V volts energy will ionize a gas molecule in 1 cm. path at 0.01 mm pressure at 28°C.[5]

C = constant in equation $P(V) = C(V - V_i)$, which holds rather accurately in the range V_i to $2V_i$. Most of the ionization by direct electron impact in gas discharges occurs within this range except in cases of very high voltage and low gas pressure.

Gas	V_i	$P(20)$	$P(30)$	$P(50)$	$P(100)$	$P(300)$	C
Cs	3.88 volts						
K	4.32						
Na	5.12						
Hg	10.39	0.067	0.146	0.195	0.213	0.193	0.00753
He	24.53	.000	.002	.009	.016	.018	.00039
Ne	21.47	.000	.003	.010	.026	.035	.00037
A	15.69	.017	.052	.093	.112	.094	.00365
*H_2	15.8±	.006	.021	.033	.038	.032	.00156
*N_2	16.3±	.004	.028	.072	.101	.101	.00223

* In the case of polyatomic gases the minimum ionizing potential is not as sharply defined as in monatomic gases, because there is a probability that the molecule will suffer a change in nuclear separation simultaneously with its ionization. This probability can be calculated by the quantum mechanics,[6] and the probability of a large change is found to be small. The experimental values given are therefore sufficiently accurate for ordinary purposes.

The impacting electron and the new electron formed in an ionizing collision may in subsequent collisions with gas molecules produce new ions. When the free paths of the electrons are small compared to the dimensions

TABLE II. *Total ionizing power of electrons.*[7] *Ions per primary electron of V volts.*

Gas	$V=30$	50	75	100	150	0
He		1.2		2.9		
Ne			1.2	2.0		
Ar	0.45	0.9		1.6	2.2	
H_2				1.4		2.8
N_2			1.3	1.6		
Hg	1.1	1.4		2.7		5.3

of the gas space, or where the boundaries are negatively charged so that the electrons continue to ionize until their energies fall below the ionizing potential, the total number of ions produced by each primary electron becomes independent of gas pressure.[7] Recently improved methods for measuring this total ionizing power of electrons of various energies have been developed

[5] Compton and Van Voorhis, Phys. Rev. **27**, 724 (1926). Jones, *ibid.* **29**, 822 (1927), who used a method for the lower voltages superior to that of Compton and Van Voorhis in insuring more uniform energies of the ionizing electrons, obtained values of P for Hg similar to those of Compton and Van Voorhis, but approximately 25 percent larger at the lower values of V. This suggests that perhaps the constant "C" should be about 25 percent larger than shown in Table I for all the gases. The values above 100 volts are probably quite accurate.

[6] Condon, Phys. Rev. **32**, 858 (1928).

[7] Langmuir and Jones, Phys. Rev. **31**, 402–3 (1928).

which yield values somewhat larger for He and smaller for other gases than were given by earlier work.[8] Some of these values are given in Table II.

If the impacting electron has more than the minimum amount of energy required to ionize the molecule, the excess of energy must be retained by the impacting electron, or transferred to the ejected electron, or used in exciting or further ionizing the ion,—or to any combination of these. From the laws of momentum transfer it is obvious that the energy transformed into kinetic energy of the ion will be insignificant. There is evidence that the excess energy is not retained wholly by the impinging electron but may be shared in any proportion between the two electrons,[9] though there is some indication that its even distribution between them is considerably less probable than the retention of most of it by one or the other.[10]

The immediate *products of such ionization* are the ejected electron and a positive ion which is simply the original molecule minus an electron.[11] In many polyatomic molecules the internal energy of this "primary" ion is greater than that of its dissociated products, so that it may split up into a positively charged and one or more neutral "secondary" products of ionization. Generally this secondary transformation requires some external stimulus, such as a collision, even though it is energetically possible without external aid, but there are a few cases known where this dissociation, subsequent to ionization, follows spontaneously and perhaps instantaneously.[12] This behavior subsequent to ionization has been well explained[13] in terms of the relation between internal energy and nuclear separation of the atomic constituents of the neutral molecule and of the primary molecular ion, which relations can be found from a study of the band spectrum. In hydrogen, for example, the primary product of ionization is H_2^+. If the gas pressure is such that these ions collide with molecules, they tend to dissociate or to associate[14] into $H^+ + H$ or $H_3^+ + H$, the relative amount of H_3^+ being greater if the effective temperature (mean kinetic energy) of the ions is low.[15] It was formerly thought that polar molecules always dissociated into positive and negative atom-ions upon ionization, because the ionizing potentials of the halogen halides could be accurately calculated from a thermodynamic cycle based upon this hypothesis.[16] This is now known to have been a fortuitous agree-

[8] Johnson, Phys. Rev. **10,** 609 (1917); Lehmann and Osgood, Proc. Roy. Soc. **A115,** 609 (1927).

[9] Eldridge, Phys. Rev. **20,** 456 (1922); Langmuir and Jones, Phys. Rev. **31,** 385–7 (1928).

[10] Harnwell, Phys. Rev. **34,** 661 (1929).

[11] Smyth, Proc. Roy. Soc. **A102,** 283 (1922); **A104,** 121 (1923); Phys. Rev. **25,** 452 (1925); Hogness et al, Phys. Rev. **26,** 44, 284 (1925); **30,** 26 (1927); **32,** 784, 936, 942 (1928).

[12] Hogness *et al,* Phys. Rev. **32,** pp. 784, 936, 942 (1928); Kallmann and Rosen, Zeits. f. Physik **58,** 52 (1929).

[13] Franck, Trans. Far. Soc. **21,** part 3 (1925); Condon, Phys. Rev. **28,** 1182 (1926); **32,** 858 (1928); Winans and Stueckelberg, Proc. Nat. Acad. Sc. **14,** 867 (1928); Condon and Smyth, *ibid.* **14,** 871 (1928).

[14] Hogness and Lunn, *ibid.* **26,** 44 (1925).

[15] Brasefield, Phys. Rev. **31,** 52 (1928).

[16] Born, Verh. d. D. Phys. Ges. **21,** 13, 679 (1919); Grimm, Zeits. f. Physik Chem. **102,** 113, 141, 504 (1922); Foote and Mohler, "Origin of Spectra," Chap. VIII (Chem. Cat. 1922).

ment. In HCl, for example, the primary product of ionization is HCl^+, and there appear to be no secondary products.[17]

(2) **Photo-ionization** may occur if the normal molecule is exposed to radiation of frequency greater than the value of ν_i in the relation $h\nu_i = eV_i$. The probability of such ionization is proportional to the radiation density and varies with frequency. Quantitative experimental studies are extremely difficult because, except in the case of the alkali metal vapors, the effective radiation lies too far in the ultraviolet to permit easy estimation of radiation intensities. With light of this spectral region, moreover, photo-electric emission of electrons from the walls of the apparatus masks the gas ionization unless special precautions are taken.

The probability B_ν of photo-ionization may be defined as the probability that an atom, exposed to unit density of radiation of frequency ν, will be ionized within unit time. All theories agree in predicting that this probability is maximum for radiation at the arc series limit, of frequency ν_i, and decreases rapidly with increasing frequency. In general the experiments confirm this. For very high frequencies (X-rays) there is the well-known Owen's law, according to which the absorption coefficient of a substance is inversely proportional to the cube of the frequency. Since the energy absorbed in any absorption act is $h\nu$, we see that Owen's law may be stated in the form: "the probability of an absorption act varies inversely as the fourth power of the frequency." Since absorption produces photo-ionization, Owen's law is equivalent to

$$B_\nu = C\nu^{-4}.$$

By the statistical principle of detailed balancing, there is derived a relation[18]

$$\frac{B_\nu}{q_\nu} = \text{const}\,\frac{\nu - \nu_i}{\nu^3} \tag{1}$$

between B_ν and the effective cross section of the atom for recombination q_ν (which is equal to the electron velocity v times the coefficient of recombination α of an ion and an electron of velocity v to form the atom). But the effective recombining cross section q_ν is not known experimentally with accuracy (see section C2) and theories regarding it are not sufficiently developed to be strictly applicable. Milne[16] assumed that the probability that an electron wil be captured by an ion varies inversely as the square of their relative velocities, *i.e.* that $q_\nu \alpha$ "varies as" $1/v^2$, which led directly to

$$B_\nu = C'\nu^{-3}$$

since $\frac{1}{2}mv^2 = h(\nu - \nu_i)$ (see Eq. 65 in section C 2). This assumption is justifiable as an approximation, for Milne shows that it is the asymptotic form of q_ν as the relative velocity approaches zero. Morse and Stueckelberg[19] have recently proved this also by wave mechanics and have, moreover,

[17] Barton, Phys. Rev. **30**, 614 (1927).

[18] Becker, Zeits. f. Physik **18**, 325 (1923); Milne, Phil. Mag. **47**, 209 (1924).

[19] Morse and Stueckelberg, to appear in Phys. Rev.

shown how q_ν varies in general with velocity for various states of the hydrogen atom. These latter expressions, while exact, have thus far proved too complicated to permit of expression in form suitable for substitution in Eq. (1). Thus even the simplest case, that of hydrogen, has resisted exact solution. Kramers[20] derived Owen's law. Becker[18] assumed that q_ν varies with velocity in the same way that the probability that a canal ray robs an electron from a neutral molecule is found to vary with their relative velocity[21] and obtained from Eq. (1)

$$B_\nu = \frac{\text{const}}{(\nu - \nu_i)\nu^3}\left[1 - \frac{(\nu - \nu_i)^2}{4\nu_i^2}\right] \tag{2}$$

which is inapplicable for $\nu \gg \nu_i$ but reduces for ν close to ν_i to

$$B_\nu = \frac{\text{const}}{(\nu - \nu_i)\nu^3}. \tag{3}$$

By methods of the recent quantum theories Suguira, Oppenheimer and Reiche have derived expressions[22] for the case of a hydrogenic atom in which B_ν varies as $1/\nu^5$, $1/\nu^{4.3}$, $1/\nu^5$ respectively. Obviously the situation on the theoretical side is far from solved with any degree of assurance.

On the expermental side, recent very careful work[23] shows conclusively that photo-ionization in vapors of calcium and rubidium agrees well with Eqs. (2 and 3), and not at all with any of the other equations. Even in these cases only relative values for different frequencies are known. In the case of potassium vapor[24] all results to date are in serious conflict with all the above theories, although this may be due to a complicating contribution of K_2 molecules to the observed ionization.

Thus, although it is observed in these vapors and some others,[25] we are as yet unable to estimate with confidence the role of photo-ionization in gas discharges. It is generally believed, however, that the absolute magnitude of photo-ionization is generally very inferior to that of ionization by electron impacts.

It should be stated that some photo-ionization has been found in mercury[26] and alkali metal vapors[23] for radiation of frequency less than the limit ν_i,—particularly for radiation belonging to the line absorption spectrum. This is undoubtedly due to cumulative effects involving two or more con-

[20] Kramers, Phil. Mag. **46,** 836 (1923).

[21] Rückhardt, Ann. d. Physik **71,** 377 (1923); Zeits. f. Physik **15,** 164 (1923).

[22] Suguira, Jour. de Phys. **8,** 113 (1927); Oppenheimer, Zeits. f. Physik **41,** 268 (1927); Reiche, *ibid.* **53,** 168 (1929).

[23] Kunz and Williams, Phys. Rev. **22,** 456 (1923); Foote, Mohler and Chenault, *ibid.* **27,** 37 (1926); Lawrence and Edlefsen, *ibid.* **34,** 233 (1929).

[24] Williamson, Proc. Nat. Acad. Sc. **14,** 793 (1928); Lawrence and Edlefsen, Phys. Rev. **34,** 1056 (1929).

[25] Mohler, Phys. Rev. **28,** 46 (1926).

[26] Rouss and Giddings, Proc. Nat. Acad. Sc. **11,** 514 (1925); **12,** 447 (1926); Houtermans, Zeits. f. Physik **41,** 619 (1927); Foote, Phys. Rev. **29,** 609 (1927).

tributions of energy. However, this effect seems to be larger than can be accounted for by any of the proposed theories.

A review of the subjects of Recombination and Photo-ionization has just been published by Mohler.[27]

(3) **Cumulative ionization** refers to ionization in two or more stages. This may be accomplished in a variety of ways. By electron impact or by light absorption a molecule may be excited to a state which falls short of complete ionization. It will then revert to a state of lower energy either spontaneously and with the emission of radiation, or as the result of a subsequent collision in which the energy may be radiated, or transferred to the impacting molecule. Experiments show that ordinarily an atom or molecule, if undisturbed by collisions, remains in the excited state for a time interval of the order of 10^{-8} sec.[28] If this state happens to be a "metastable" state, from which spontaneous escape by radiation is inhibited according to the "selection principles" of the quantum theory, the lifetime in this state is much longer. If the gas pressure is high so that molecular collisions are frequent, the life may be limited to the average interval before the next collision. But in any situation an excited molecule has a certain average period of existence denoted by τ. If, while in this excited state, it is struck by another electron or absorbs more radiant energy, or collides with another excited molecule, it may acquire enough additional energy to complete its ionization. *Successive impact, successive absorption*, or a combination of an absorption and an impact which may be termed *photo-impact* etc. denote the various ways in which this cumulative ionization may occur.

The rate of cumulative ionization is obviously jointly proportional to the intensities of the contributing ionizing factors, and to the average lifetime τ of the excited state. If the ionization is due entirely to current through the gas, the rate of two stage cumulative ionization should be proportional to the square of the current, other factors being assumed constant.

Except at the very lowest pressures, there are two phenomena which enhance the rate of cumulative ionization by increasing the concentration of excited molecules. The first of these is the strong absorption by the gas of the *"resonance" radiation* which arises when molecules revert from an excited state to the normal state. By this absorption other molecules are in turn excited,—and they in their turn will excite still others when they radiate their energy. In this way the energy is passed on from molecule to molecule, thus multiplying the chance that each quantum of energy may ultimately contribute to ionization. This passage of resonance radiation through a gas can be treated as a problem in diffusion[29] by the equation

[27] Mohler, Phys. Rev. Supp. **1,** 216 (1929).

[28] Wien, Ann. d. Physik **60,** 597 (1919); **73,** 483 (1924).

[29] Compton, Phil. Mag. **45,** 570 (1923); Milne, Jour. London Math. Soc. **1,** 1 (1926); Zemansky, Phys. Rev. **29,** 513 (1927). Milne's equation is more general than Compton's Eq. (4) in that it is not limited in application to cases in which the dimensions of the apparatus greatly exceed the effective "free path" of the radiation.

$$\int\int \frac{1}{3}\,\frac{1}{\alpha^2\tau}\,\frac{dN'}{dn}\,dS = -\int\int\int R\,dxdydz \qquad (4)$$

where α is the absorption coefficient for the radiation, τ the average life of an excited state, N' the concentration of excited molecules, n the normal to a surface element dS and R the total rate of production of new quanta of resonance radiation within the closed surface. The second phenomenon is the tendency of molecules in excited states to go into *long lived metastable states* as the result of collisions with other molecules. It is still uncertain as to which of these phenomena is the more important in the cumulative ionization of mercury vapor and the noble gases. In the vapors of the alkali metals, which have no metastable states, the first process alone can be effective.

TABLE III. *Average life times τ of some excited molecules.*[30]

Molecule	Excited to emit:	Average life
H	H_α H_β or H_γ	$1.85(10)^{-8}$ sec.
Hg	$1s^1S_1-2p^3P_1(2536)$	9.7
Hg	$2p^3P_1-2s^3S_1(4358)$	1.81
Na	$1s^2S_1-2p^3P_{1,2}(NaD_{1,2})$	3.70
N	arc lines	9.33
N	spark lines	1.35

The experimental values of Wien's in Table III are not necessarily the true values of τ characteristic of the excited states in question, since the conditions of the experimental measurements permitted also higher states of excitation, from which the lower states in question could be replenished in one or more stages. For example hydrogen atoms excited to the fifth quantum state can emit H_γ by reversion directly to the second quantum state. But they may also revert successively to the fourth, third and second states, emitting H_α in the last transition. In this case the value of τ measured for H_α would be the sum of τ's for the three transitions. Wien's method gives a value of τ which is a weighted mean of all such processes. The probabilities of the transitions are such, however, that the experimental values are not greatly in error from such complications. Methods which are free from this complication can be used to determine the life times of some optically excited states.[31]

Maxwell[32] has measured τ for lines of mercury atoms in various states of ionization and found values ranging between 10^{-8} and $9(10)^{-7}$ sec. for lines of Hg II, Hg III, Hg IV, with a regular gradation of values which suggests that the mean life is longer for higher states of ionization. His calculation of the mean life τ of an excited state from the aggregate rate of its depopulation leads to

$$\tau = \left(\sum_{n',k'} A_{n,k}^{n',k'} \right)^{-1}$$

[30] Wien, Ann. d. Physik **60**, 597 (1919); **73**, 483 (1924); Kerschbaum, *ibid.* **79**, 465 (1926); Dempster, Astrophys. Jour. **57**, 193 (1923).

[31] Cario, Zeits. f. Physik **10**, 185 (1922); Turner, Phys. Rev. **23**, 464 (1924).

[32] Maxwell, Phys. Rev. **84**, 199 (1929).

where the summation is taken over all the states n', k' to which the state may change, $A^{n',k'}_{n,k}$ being the transition probability of such change. This expression, previously given by Tolman[33] and others, indicates that the same life should be found from measurements of all spectral lines originating from the same state. This expression also neglects the repopulation of the state in question from higher states.

The average "natural" life times of metastable states are of a larger order of magnitude than those of other excited states. Saha and Kothari[34] derived the theoretical expression

$$\tau = \frac{3c^5m^2}{\pi^2e^2h\nu^3}$$

for the natural lifetime of a metastable atom whose "forbidden" line has wave number ν. For $\lambda = 10{,}000\ A$ this leads to $\tau = 0.15$ sec.

In actual experimental discharge tubes the lives of the metastable atoms are limited by collisions with the walls of the apparatus or with other molecules, so that the actual lifetime is a characteristic of the apparatus rather than of the atom itself. Under such conditions it is given by a relation

$$\frac{1}{\tau} = \frac{A}{p} + Bp$$

where p is the pressure and A and B involve effective collision radii, dimensions of apparatus, etc. The term A/p gives the rate of disappearance of metastable atoms at the walls of the apparatus and the term Bp gives the rate of disappearance within the apparatus. Couliette[36] showed that metastable atoms of Hg diffuse through Hg according to the ordinary laws of gas diffusion, and as if the effective radius of a metastable atom is 1.5 times that of a normal atom. Much work has been done to discover the probability that a metastable atom will revert to the normal state as the result of collisions with other atoms or molecules. Apparently the most effective destroying collisions are those with molecules which can absorb the available energy by becoming themselves excited or dissociated,—such as dissociation of H_2 by metastable Hg. Dorgelo[35] found metastable atoms of neon in neon to persist as long as 0.10 sec. A critical survey of recent work in this field,[36] with an interesting theoretical analysis, has just been published by Zemansky.[37]

The *concentration of metastable atoms* in a discharge tube is very important to know, if they are active agents in ionization either of the gas or the electrodes. The only methods suitable for measuring this concentration are

[33] Tolman, Phys. Rev. **23,** 693 (1924).

[34] Saha and Kothari, Naturwiss. **17,** 271 (1929).

[35] Dorgelo, Physica, **5,** 429 (1925).

[36] Metastable neon in neon, Meissner and Graffunder, Ann. d. Physik **84,** 1009 (1927); in H_2, N_2, A, Eckstein, *ibid.* **87,** 1003 (1928). Metastable Hg in Hg, Zemansky, Phys. Rev. **29,** 513 (1927); Coulliette, *ibid.* **32,** 636 (1928); in N_2, Pool, Phys. Rev. **33,** 22(1929); Asada, Phys. Zeits. **29,** 708 (1928).

[37] Zemansky, Phys. Rev. **34,** 213 (1929).

optical ones, based on absorption or dispersion,[38] and these require extreme care in technique and interpretation. For neon in the positive column of glow discharges, Kopfermann and Ladenburg found the concentrations in the metastable states s_3, s_4 and s_5 to be given approximately by

$$N_s = \frac{aJ}{bJ+c} \tag{1.5}$$

where J is the current density and a, b, c are production and extinction factors which are independent of current but depend on pressure. For the state s_5 they give $a = 3.5(10)^{11}$, $b = 0.365$, $c = 1$ with J in milliamperes per cm^2. It is thus obvious that the concentration of metastable atoms may be as large as $(10)^{12}$, and therefore comparable with that of positive ions (as we shall see in Part II).

(4) **Ionization by positive ion impact** is well known in the cases of α particles and canal rays and has long been assumed[39] to occur to some extent with positive ions of much smaller velocity. Ions apparently begin to be very effective ionizing agents when their velocities are as large as those of electrons which have fallen through the minimum ionizing potential,[40] which means positive ions of energies corresponding to tens of thousands of volts. When their energies are much smaller, however, they become exceedingly ineffective, so that the ionization due to these impacts is masked by a variety of complicating secondary ionizing effects which are more or less inherent in any method of experimental investigation. As a result, our direct knowledge of this phenomenon is very meager and much of the evidence is contradictory.[41]

In apparatus which could have detected about one ionizing act for every thousand primary positive ions, no evidence of ionization of hydrogen by sodium or potassium ions was found for velocities of the ions up to about 1000 equivalent volts.[42] In the noble gases, ionization by potassium ions could be detected at velocities as low as 100 volts, and the trend of values in this region indicated that the probability of ionization would fall, if not strictly to zero, at least to very much less than 0.1 percent at velocities somewhat below 100 volts. The ionization of several gases by potassium ions at 750 volts, expressed as the number of ions formed per initial positive ion per cm path at 0.01 mm pressure, is:[43] argon 0.00288; neon 0.00112; nitrogen 0.00124; air, 0.00098; hydrogen 0.0000. Perhaps the most nearly direct evidence of ionization at velocities as low as 40 volts is found in the excitation of mercury arc lines in mercury vapor bombarded by sodium ions,[44] but quantitative

[38] De Groot, Zeits. f. Physik **55**, 52 (1929); Kopfermann and Ladenburg, *ibid.* **48**, 15, 26, 51 (1928); excellent discussion of the general theory of dispersion.

[39] Townsend, Electrician, April 3 (1903); "Electricity in Gases," Chap. IX (Oxford 1915).

[40] J. J. Thomson, "Rays of Positive Electricity," 1 ed. p. 36.

[41] Loeb, Science **66**, 627 (1927).

[42] Hooper, Jour. Frank. Inst. **201**, 311 (1926).

[43] Sutton, Phys. Rev. **33**, 364 (1929); Sutton and Mouzon, *ibid.* (in print).

[44] Tate, Phys. Rev. **23**, 293 (1924).

conclusions cannot be drawn from this. Experimental difficulties have thus far made it impossible to draw any definite conclusions regarding the efficiency of ionization of molecules of a gas *by ions of the same gas*, except for showing that this efficiency is of a smaller order of magnitude than that of ionization by electron impacts. Direct experiments designed to discover the ionizing efficiencies of positive ions of energies in the range of those to be expected in arcs, sparks and glow discharges are very much needed.

(5) **Ionization by collisions of the second kind** is ionization effected by a transfer of energy from an excited or ionized molecule to some other molecule. The prime requisite is that the energy available be adequate for the ionization. The probability of this process seems always to be maximum if the available energy is just sufficient, and to decrease with excess of energy. Thus, at contact, an ion of one gas may capture a neutralizing electron from a neutral molecule of another gas of lower ionizing potential,[45] thereby ionizing the latter, and at the same time the latter ion may also be raised to an excited state.[46] In these cases one type of ion is simply exchanged for another. But new ions also may be produced, as by the action of metastable helium atoms in the 2^3S or 2^1S states (of 19.77 or 20.55 volts energy) in ionizing atoms of impurities at contact (all requiring less energy).[47]

(6) **Thermal ionization** of a gas may be treated as a case of thermal dissociation in which the products are an electron and a positive ion, and the degree of ionization may be calculated as a function of the temperature T and pressure P of the gas by applying Nernst's heat theorem. This idea has been applied with striking success to problems of ionization in stellar atmospheres by use of the so-called "Saha" equation and its successive refinements.[48] There have been attempts to apply this equation to calculate the degree of ionization of gas in an enclosure,[49] but these are seriously wrong because of neglect of the very important part which may be played by emission of electrons from the walls of the enclosure. But a more general equation will apply equally well to a star or to a laboratory device, *viz.*

$$\log_{10}\frac{n_e n_p}{n_a} = -\frac{5040 V_i}{T} + \frac{3}{2}\log_{10} T + 15.385 \qquad (5)$$

where n_e, n_p and n_a are the numbers of electrons, positive ions and atoms, respectively, per cm^3; V_i is the ionizing potential of the gas in volts and

[45] Harnwell, Phys. Rev. **29**, 683, 830 (1927); Hogness and Lunn, *ibid.* **30**, 26 (1927); Smyth and Stueckelberg, *ibid.* **32**, 777 (1928).

[46] Duffendack and Smith, Nature, May 21, 1927; Duffendack, Henshaw and Goyer, Phys. Rev. **34**, 1132 (1929).

[47] Franck and Jordan, "Anregung von Quantensprüngen durch Stösse," p. 119 (Springer 1926). See also Penning, Zeits. f. Physik **57**, 723 (1929, for the ionization of argon by metastable neon.

[48] Saha, Phil. Mag. **40**, 472 (1920); Proc. Roy. Soc. **99**, 135 (1921); Tolman, Jour. Am. Chem. Soc. **43**, 1630 (1921); Russell, Astrophys. Jour. **55**, 119 (1922); Fowler and Milne, Monthly Notices Roy. Astr. Soc. **83**, 403 (1923); Milne, *ibid.* **85**, 111 (1924); **86**, 8 (1925).

[49] Saha, Phil. Mag. **46**, 534 (1923).

T is the absolute temperature.[50] In particular, Langmuir and Kingdon showed that the proportional ionization of caesium atoms incident upon a hot metal surface is a thermal ionization which may accurately be calculated by this equation.

B. Liberation of Electrons and Ions from Electrodes

All processes of emission from electrodes are greatly dependent on their surface condition as regards surface films of foreign metals or adsorbed gases. For this reason it is not certain how accurately experimental data from special studies of these processes can be carried over to the actual conditions in gas discharges. It is likely that this can be done at least to a first approximation, though it is desirable, from the standpoint of the theory of gas discharges to investigate as many of these processes as possible under the conditions which obtain in the discharge.

(1) **Thermionic emission of electrons** is capable of yielding currents limited only by the temperatures to which the cathodes can be raised. At the melting point of tungsten, for example, the saturation emission is 480 amp./cm^2. The saturation current density j_s is a function only of the temperature T and the nature of the metal as defined by a certain constant ϕ_0 known as its "*work function*" and by another constant A:

$$j_s = AT^2\epsilon^{-e\phi_0/kT} = AT^2\epsilon^{-b/T}, \tag{6}$$

where, by putting $e/k = 11{,}606$ degrees per volt, we can express ϕ_0 in equivalent volts. This equation was originally derived by a thermodynamical argument[51] based on the assumptions (1) that the potential energies of an electron inside and outside the metal surface differ by $e\phi$; (2) that $e\phi = e\phi_0 + (3/2)kT$, where ϕ_0 is independent of temperature; (3) that the electrons outside the metal follow the gas law $pv = nkT$; (4) that all electrons striking a surface are absorbed by it. Of these, (2) results as an approximation from a consideration of the Thomson coefficient based on assumptions (1) and (3), (3) is based on the experimental fact that the emitted electrons possess a Maxwellian distribution of velocities characteristic of the temperature of the emitting metal,[52] and (4) is known to be only approximately true, but the error thus introduced is probably nearly independent of temperature T(see ref. 57 below). Direct experimental confirmation of the relation

$$e\phi = e\phi_0 + (3/2)kT \tag{7}$$

is very difficult. Such evidence as is available suggests a relation of this sort but with a somewhat larger temperature coefficient.[53]

[50] Langmuir and Kingdon, Proc. Roy. Soc. **107**, 61 (1925); Fowler, "Statistical Mechanics," p. 281 (Cambridge 1929).

[51] Richardson, "Emission of Electricity from Hot Bodies," Chap. II (1916); Phil. Mag. **23**, 601, 619 (1912); *ibid.* **24**, 740 (1912); *ibid.* **28**, 633 (1914).

[52] Richardson and Brown, Phil. Mag. **16**, 353 (1908); (see also ref. 57 below).

[53] Davisson and Germer, Phys. Rev. **20**, 300 (1922).

Subject to neglect of the contribution of electrons to the specific heat of a metal, to neglect of energy of surface charge, to acceptance of the Sakur-Tetrode chemical constant as the appropriate constant of integration for the Clapeyron equation and to assumption of zero reflecting power of all metals for electrons, it is predicted that A should be a universal constant[54] of value $A_0 = 60.2$ amp./cm^2 deg^2, and this seems to be the right value for a number of metals rather similar to tungsten. There is some evidence, however, that A is itself a function[55] of ϕ_0, such as $A = A_0 \exp(e\alpha/k)$, in which α is the negative temperature derivative of ϕ_0. For tungsten $\alpha = 0$, for platinum α is positive and for thoriated tungsten it is negative.

On the basis of Sommerfeld's recent theory of metals,[56] he and Fowler[57] have again derived Eq. (6), but with certain differences in interpretation. Fowler finds $A = 2A_0(1-r)$, where the factor 2 enters because of the two possible orientations of the electron spin vector, and r is the fraction of the electrons, incident on the metal surface from a surrounding electron atmosphere, which are reflected. On this theory, the variation in A for different metals is due to characteristic differences in this reflecting power r (a fact also recognized by Richardson). The experiments would support this theory if it could be shown that $r = 1/2$ for tungsten (for which $A = A_0$), $r < 1/2$ for platinum and $r > 1/2$ for thoriated tungsten. There is some little experimental suggestion that $r \sim 1/2$ for very slow electrons, but these experiments are subject to corrections which become very uncertain at low velocities. There are furthermore, convincing theoretical reasons against assuming an appreciable value of r for very slow electrons (see B 7). Thus Fowler's theory cannot be said definitely to be supported by experiments on this point. Sommerfeld shows that the work function $e\phi$ is the difference between the work required to take an electron out of the metal against the attractive forces at the surface and the contribution of the pressure of the conducting electrons within the metal toward overcoming these forces.

Despite these uncertainties in interpretation, Eq. (6) is one of the most accurately verified equations in the whole field of physics, since it holds within the degree of accuracy with which temperatures can be measured throughout a current range of about a million-million fold. Furthermore there is no doubt about the essential correctness of the interpretation of ϕ as the work necessary to remove a unit charge of electrons from a metal.

The more electropositive metals are characterized by smaller work functions and hence emit larger thermionic currents at any given temperature than do the electronegative metals. They are not, however, generally suitable for use as high temperature cathodes on account of their volatility. However, monatomic films of such substances on more electronegative metals adhere

[54] Dushman, Phys. Rev. **21**, 623 (1923); see also Richardson, "Emission of Electricity from Hot Bodies," p. 39.

[55] Du Bridge, Proc. Nat. Acad. Sc. **14**, 788 (1928); Bridgman, Phys. Rev. **31**, pp. 90, 862 (1928).

[56] Sommerfeld, Zeits. f. Physik **47**, 1 (1928).

[57] Fowler, Proc. Roy. Soc. **A117**, 549 (1928); *ibid.* **122**, 36 (1929).

with remarkable tenacity and constitute convenient sources of low temperature emission. The best known of these are thorium[58] and caesium[59] on tungsten, which are typical of two methods of coating.

Thorium films are produced by diffusion of metallic thorium toward the surface. This diffusion is particularly rapid along crystal grain boundaries but must occur also from the interior of crystals to these boundaries and probably also through the surface layer of crystals. This diffusion occurs at a convenient rate at about 2100°K, and proceeds until the surface is covered by a monatomic layer of thorium. Such surfaces can be operated to give thermionic emission of about 1 amp. per cm^2 at the maximum temperature at which thorium is not lost by evaporation faster than it is replaced by diffusion.

Caesium films are produced by the deposit of caesium atoms from its vapor upon the tungsten surface. Its adherence up to temperatures as high as 900°K depends upon the fact that the ionizing potential of caesium atoms is less than the work function of tungsten, so that the tungsten robs the caesium atoms of their valence electrons when they strike the surface, and holds them on the surface as ions. If the temperature is raised high enough to evaporate them off the surface, they leave as ions so long as the work function of the surface exceeds the ionizing potential of the atom. If a metal of larger work function than tungsten is used, such as oxidized tungsten, the caesium adheres still more firmly and the maximum temperature at which emission can be obtained without loss of the film is raised. Similar effects are observed with vapors of rubidium and potassium.[60]

The degree of activation of such coated surfaces is conveniently expressed by a quantity θ, called by Langmuir "the fraction of the surface covered by adsorbed ions." This is approximately a linear function, of the work function so that

$$\theta = (\phi'' - \phi)/(\phi'' - \phi'), \tag{7.5}$$

where ϕ, ϕ' and ϕ'' are the work functions of the actual surface, of the most completely (effectively) coated surface, and of the clean surface, respectively.[58,59]

In gases at low pressures these activated surfaces, such as those of thorium, are easily destroyed (knocked off) by positive ion bombardment when the anode voltage is raised above certain critical values.[61]

The distribution of velocities of thermally emitted electrons is Maxwellian, and characteristic of the temperature of the emitting surface.[62] The mean kin-

[58] Langmuir, Phys. Rev. **22,** 357 (1923).

[59] Langmuir and Kingdon, Science **57,** 58 (1923); Proc. Roy. Soc. **107,** 61 (1925); Becker, Phys. Rev. **28,** 341 (1926); **29,** 364 (1927); Bell Telephone Laboratories Reprint No. B-412, August, 1929.

[60] Killian, Phys. Rev. **27,** 578 (1926).

[61] Kingdon and Langmuir, Phys. Rev. **22,** 148 (1923); A. W. Hull, Transaction Amer. Inst. of Electrical Engs. **49,** 753 (1928).

[62] Richardson and Brown, Phil. Mag. **16,** 353 (1908); Schottky, Ann. d. Physik **44,** 1101 (1914); Jones, Proc. Roy. Soc. **A102,** 734 (1923); Germer, Phys. Rev. **25,** 795 (1925).

TABLE IV. *Some thermionic work functions.*[63]

Metal	ϕ_0 (volts)	Metal	ϕ_0 (volts)
Hf	5.09	Th	3.35
*Pt	4.8 –6.35	**Th	2.63
W	4.52	**U	2.84
Mo	4.41	**Ce	2.72
Ta	4.07	**La	2.72
Zr	4.50	***BaO	1.68–3.44
**Zr	3.15	BaO+SrO	1.51–1.89
Cs	1.81	CaO	2.19
**Cs	0.72	MgO	1.02

* Pt, perhaps because of its electronegative character, is extremely hard to free from hydrogen and electropositive contaminants. It is only after the severest purification and degassing that the characteristic value 6.35 is reached.

** Monatomic layers on tungsten.

*** Values depend on treatment of the surface (flashing, bombardment, etc.).

etic energy of the electrons in equilibrium in a given volume is $(3/2)kT$, being $(1/2)kT$ for each of the three dimensions. For the electrons which cross or are emitted from a surface the mean kinetic energy is still $(1/2)kT$ for each of the two dimensions parallel to the surface, but the mean energy corresponding to the velocity component normal to the surface is kT, making a total of $2kT$ for the mean energy of the emitted electrons.

These energies may readily be expressed in equivalent volts defined by the relation

$$Ve = kT. \tag{8}$$

The value of e/k is conveniently expressed[64] as 11,606 degrees per volt $(10^8 \times 1.5911 \times 10^{-20}/1.3709 \times 10^{-16})$. Thus the mean energy $(2kT)$ of the electrons emitted at temperature T corresponds to $\overline{V}$ given by

$$\overline{V} = 2kT/e = 2T/11{,}606 \text{ volts}. \tag{9}$$

This mean energy corresponds to only 0.052 volt at room temperature 300°K, 0.41 volt at 2400°K and 0.63 volt at 3655°K, the melting point of tungsten.

The fraction n/n_0 of the emitted electrons which are capable of moving against a retarding field of V volts can be calculated from the Boltzmann equation[65]

$$n/n_0 = \epsilon^{-Ve/kT}. \tag{10}$$

Thus, for example, only one out of a thousand electrons from a hot cathode at 2400°K can move into regions which are more than 1.43 volts negative with respect to the cathode.

The heat of evaporation of electrons from a metal surface is the heat absorbed in the process of emitting an electron. It is analogous to the latent heat of evaporation of molecules from a liquid surface and is due to the fact

[63] Dushman, *ref.* 54; Dushman, Dennison and Reynolds, Phys. Rev. **29,** 903 (1927); Du Bridge, Phys. Rev. **31,** 236 (1928); Dushman, Int. Critical Tables VI, p. 53.

[64] Birge, Phys. Rev. Supp. **1,** 1 (1929).

[65] Langmuir and Mott-Smith,Jr., G. E. Rev. **27,** 449 (1929); Mott-Smith and Langmuir, Phys. Rev. **28,** 756 (1926).

that only the fastest electrons in the metal are able to escape. It and its converse, the heat of condensation, have been measured for electrodes in high vacuum,[66] and the latter also for electrons received by the anodes in gas discharges.[67] This heat of evaporation is found to be identical with the work function $e\phi$. The heat of condensation is perhaps the best way to determine the work function of an electrode in a neutral or ionized gas, since it may be measured without highly heating or otherwise influencing the electrode. With saturation emission the cooling effect due to loss of electrons increases exponentially with the temperature whereas the cooling due to conduction and radiation varies roughly as the first and fourth powers of the temperature, respectively. Hence at high temperatures the cooling effect due to thermionic emission may exceed that due to other causes.

(2) **Thermionic emission of positive ions** consists usually in the emission of charged atoms of electropositive substances such as the alkali or alkaline earth metals which are present, either purposely or accidentally as impurities on the surfaces of more electronegative metals.[68] Alkali atoms do not evaporate as ions from surfaces of their own metal. They do evaporate as ions from surfaces of platinum, iron oxide, etc. It seems likely that here, as in the case of alkali metal films on tungsten,[59,60] the criterion for evaporation as an ion is very nearly that the work function of the surface exceeds the ionizing potential of the atom. When the surface conditions remain constant, this positive ion emission varies with temperature according to an equation similar to Eq. (6).[69]

Quite recently another type of positive ion emission has been found, in which the ions are charged atoms of the heated metal itself, and not impurities. This is proved by measurement of their masses by a mass spectrograph. Wahlin[70] first reported this type of emission and has found that these characteristic ions are emitted by chromium, molybdenum, tungsten, rhodium, ruthenium, tantalum and columbium. They were not detected from iron, nickel, cobalt, copper, silver, gold, iridium, platinum, zirconium, palladium or antimony. Manganese is doubtful. All these metals gave the usual alkali impurities when first heated. The first group only gave the characteristic ion emission which persisted after the impurities had disappeared with continued heating.

L. P. Smith has investigated more extensively tungsten and molybdenum.[71] He finds these currents to vary with temperature according to an equation of the general type of Eq. (6), from which he calculates positive ion work functions of $\phi_{+0} = 6.55$ volts for tungsten and 6.09 volts for molybdenum.

[65] Richardson and Cooke, Phil. Mag. **20,** 173 (1910); *ibid.* **21,** 404 (1911); Cooke and Richardson, *ibid.* 25, 624 (1913); Wehnelt and Jentzsch. Verh. d. D. Phys. Ges. **10,** 610 (1908)

[67] Schottky and Issendorff, Zeits. f. Physik **26,** 85 (1924); Van Voorhis, Phys. Rev. **30,** 318 (1927).

[68] Richardson, Phil. Mag. **16,** 740 (1908); Proc. Roy. Soc. **89,** 507 (1914); Davisson, Phil. Mag. **23,** 121 (1912); Barton, Harnwell and Kunsman, Phys. Rev. **27,** 739 (1926).

[69] Richardson, Phil. Mag. 4, 98 (1902); Kunsman, Proc. Nat. Acad. Sc. **12,** 659 (1926).

[70] Wahlin, Phys. Rev. **34,** 164 (1929) and unpublished data.

[71] L. P. Smith, Phys. Rev. **34,** 1496 (1929); **35,** 381 (1930).

These values are much smaller than other estimates based on a simple thermodynamical energy cycle. Also the term corresponding to A in the equation corresponding to Eq. (6) does not agree with its theoretical value unless these positive ions are assumed to have a reflection probability of about 0.99999 if they strike the metal surface to condense upon it!! Hence it is obvious that the physical interpretation of this interesting type of emission is far from understood.

It should be emphasized that these "characteristic ion" currents are extremely small in comparison with other thermionic emission currents. For example measurements made in the Research Laboratory of the General Electric Company gave, from tungsten, about 5.4 $(10)^{-10}$ amp. per cm^2 at 2500°K and about 1.4 $(10)^{-8}$ amp. per cm^2 at 2800°K. These values indicate that the rate of emission of these ions is far less than the rate of evaporation of neutral atoms. Thus there is only about one ion for every 2000 atoms at 2500°K and about one for every 4200 atoms evaporating at 2800°K. One might suspect that this represents thermal ionization of the emitted metal vapor, but application of Eq. (5) shows that such thermal ionization would account for ionization of only about one in every million evaporated atoms, and is therefore quite inadequate as an explanation. Another suggestion is that such ion emission is causally related to the ionic nature of the metal lattice, but this is mere speculation at present.

For many purposes an extremely useful source of positive ions is obtained from a hot filament of electronegative metal in the presence of a vapor whose ionization potential is less than the work function of the filament.[59] For example, a tungsten filament at any temperature above 1200°K will send off as a positive ion *every* caesium atom which strikes it. In this way intense sources of positive ions can be obtained even though the vapor pressure of the source of these ions is so small that the vapor plays no detectable role in the discharge, except to supply these positive ions by contact with the filament. Such a source gives a current which is independent of the temperature of the filament, but is proportional to the pressure of the vapor.[72]

(3) **Photo-electric emission of electrons** occurs if a metal is illuminated by radiation of frequency ν greater than a threshold value ν_0 given by

$$h\nu_0 = e\phi \tag{11}$$

where h is Planck's constant and $e\phi$ is the work done in removing an electron from the metal. ϕ is generally called the "photo-electric work function" of the metal. The number of electrons emitted per second is directly proportional to the intensity of the incident radiation, but the velocities of emission are independent of the intensity.

Velocities of emission of photoelectrons are distributed between zero and a definite maximum which depends only upon the frequency of the radiation and the work function of the metal, as given by Einstein's well-known photo-electric equation[73]

[72] Bainbridge, Phys. Rev. **34**, 752 (1929).
[73] Einstein, Ann. d. Physik **17**, 145 (1905).

$$\tfrac{1}{2}mv_{\max}^2 = eV = h\nu - e\phi \qquad (12)$$

which has been verified over the entire spectrum from near infrared to X-rays.[74] In it, V denotes the smallest retarding difference of potential which prevents the fastest electrons from going from the emitting to the receiving electrode. The ordinary interpretation of this equation is that the electron acquires energy $h\nu$ from the radiation and loses energy $e\phi$ in escaping.

Between zero and this maximum velocity there is a distribution of velocities such that the most probable energy is about half the maximum energy[75] in the case of ordinary or ultraviolet light acting on metal surfaces. When the photo-electric emission is caused by X-rays, the larger portion of emitted electrons have energies close to the theoretical maximum. The great generality of the $h\nu$ = energy relation suggests that those electrons which escape with less than the maximum energy must have lost the difference of energy at atomic encounters within the metal in the interval between acquiring the energy and escaping from the surface.[76]

Table V gives some typical data on initial energies of electrons, expressed in equivalent volts.

TABLE V. *Maximum emission energies of photoelectrons for various Metals and Wave-lengths.*

Metal	Threshold Å	V (volts)				
		4000	3000	2500	2000	1000
Pt[77] (outgassed)	1945	—	—	—	—	6.00
Pt[75] (ordinary)	2880	—	—	0.66	1.89	8.07
W[78] (outgassed)	2575	—	—	0.14	1.37	7.55
Hg[79] (liquid)	2735	—	—	0.42	1.65	7.83
Zn[75] (ordinary)	3760	—	0.83	1.65	2.88	9.06
K[80] (clean)	7000	1.32	2.35	3.17	4.40	10.58
K[80] (sensitized hydride)	10000	1.85	2.88	3.70	4.93	11.11

On the basis of apparent discrepancies between contact difference of potential and difference between work functions of two metals Millikan at one time held that thermionic electrons originate as *free* or conducting electrons in a metal and that photo-electrons are originally *bound* electrons in the metallic atoms.[81] This view was supported by the fact that temperature has very little effect on photo-electric emission[82] (unless the physical condition of the

[74] Richardson and Compton, Phil. Mag. **24,** 575 (1912); Millikan, Phys. Rev. **7,** 355 (1916); Duane and Hunt, *ibid.* **6,** 166 (1915); Webster *ibid.* **7,** 599 (1916).

[75] Richardson and Compton, Phil. Mag. **24,** 575 (1912).

[76] Richardson, Proc. Roy. Soc. **A94,** 269 (1918).

[77] Du Bridge, Phys. Rev. **31,** 236 (1928).

[78] Warner, Proc. Nat. Acad. Sc. **13,** 56 (1927).

[79] Kazda, Phys. Rev. **26,** 643 (1925).

[80] Richardson and Young, Proc. Roy. Soc. **A107,** 377 (1925).

[81] Millikan, Phys. Rev. **7,** 18 (1916); *ibid.* **18,** 236 (1921).

[82] Millikan and Winchester, Phys. Rev. **24,** 16 (1907); Burt, *ibid.* **24,** 207 (1924); Koppius, *ibid.* **18,** 443 (1921); Nielson, *ibid.* **25,** 30 (1925).

surface is affected by temperature changes[83]). Later the above discrepancies were found to be due to complications in the interpretation of the contact difference of potential measurements, and the original conclusion of Richardson and Compton,[75] in regard to the identity of contact difference of potential and difference of work functions of *pure* metals, has been well confirmed. (Complications, which are discussed later, cause departures from this identity in the case of metals with composite surfaces.) Furthermore, Sommerfeld's theory of metals[56] explains the independence of photo-electric effect with temperature. Thus there is no present reason for not supposing that the same group of electrons in the metal contribute to thermionic and photo-electric emission.

It it true that photoelectrons have been observed, which come from the deep interior of atoms as "bound" electrons under the influence of X-rays or γ-rays. For these the values of ϕ in Eq. (12) include the binding energies of the electrons in the various "shells" in the atoms, and indeed Eq. (12) gives a good method of measuring these energies.[84] It is doubtful, however, whether such energies play any appreciable role in the phenomena of gas discharge.

For a time it was thought that evidence had been found for "cumulative" photo-electric emission[85] in that, at frequencies less than the ordinary threshold ν_0, there was feeble emission in which the maximum velocity appeared to be given by Einstein's equation (12) with $2h\nu$ instead of $h\nu$, but this is now found to be due to the remarkable effect of even small accelerating fields in decreasing the apparent work function of activated surfaces, as will be discussed further in the following section (B 4).

The total photo-electric emission, besides being proportional to the light intensity, is a function of the metal and of the frequency. The more electropositive metals are the more sensitive. The sensitivity appears in general to increase, with increasing frequency of radiation, from zero at ν_0 to a maximum at $\nu = 3\nu_0/2$ after which it decreases, but rises to one or more maxima at still shorter wave-lengths.[86] Any maximum in the light absorption for a given frequency would give rise to a maximum in the photo-electric emission, but at least the first maximum seems not to be thus explained. The only theory of total emission as a function of frequency, which is by Richardson,[87] seems to be dimensionally correct and predicts the maximum of sensitivity at $3\nu_0/2$, but it has not been proved to be correct in detail.

The complete photo-electric emission under the influence of black body radiation has been shown theoretically[88] and experimentally[89] to vary with the

[83] Ives and Johnsrud, Jour. Opt. Soc. Am. and R. S. I. **11,** 565 (1925).

[84] De Broglie, Jour. d. Phys. et le Rad. **2,** 265 (1921); Robinson, Phil. Mag. **50,** 241 (1925); Ellis, Proc. Camb. Phil. Soc. **22,** 369 (1924).

[85] Nottingham, Phys. Rev. **33,** 633 (1929).

[86] Compton and Richardson, Phil. Mag. **26,** 549 (1913); Richardson and Rogers *ibid.* **29,** 618 (1915); Richardson and Young, Proc. Roy. Soc. **107,** 377 (1925); Ives, Jour. Opt. Soc. Am. and R. S. I. **8,** 551 (1924).

[87] Richardson, Phil. Mag. **23,** 619 (1912).

[88] Richardson, Phil. Mag. **23,** 619 (1912); **31,** 149 (1916); **47,** 975 (1924); Wilson, Ann. d. Physik **42,** 1154 (1913).

[89] Wilson, Proc. Roy. Soc. **A93,** 359 (1917); Roy. Soc. Proc. **A112,** 599 (1926).

black body temperature according to an equation similar to Eq. (6) for thermionic emission. This suggested the attractive hypothesis that thermionic emission is simply the photo-electric emission due to the radiation density within the hot metal, integrated over the complete spectrum and throughout the body of the metal. Quantitative test is very difficult, owing perhaps principally to uncertainty regarding the depth within the surface from which photoelectrons can emerge. The best estimates place the integrated photo-electric emission at about 5000 times less than the thermionic emission,[90]—this difference being attributed to the factor A of Eq. (6).

Experiments on the photo-electric effect from thin films of metal on quartz have shown that photo-electrons may escape from depths of several atomic diameters within the surface of the metal.[91] The average distance which an electron may move in the metal without losing its ability to escape appears to be about 2.7 $(10)^{-7}$ cm in platinum and 5.0 $(10)^{-7}$ cm in gold, and the probability of going a given distance without losing ability to escape falls off exponentially with the distance.

Finally, mention must be made of the *selective photo-electric effect*[92] which is found with electropositive metals and usually when the electric vector in the incident radiation has a component normal to the surface of the metal. It appears as an increased sensitivity reaching a maximum at the same frequency $\nu = 3\nu_0/2$ discussed above as the region of maximum photo-electric sensitivity. It was formerly thought to depend on some different mechanism from that of the *normal* photo-electric effect, but it is now believed to be identical in nature, and to exist in virtue of selective absorption in this spectral region and with light polarized so as to have its electric vector normal to the surface.[93] This theory has, however, not been entirely adequate to explain the observations.[94]

(4) **Relations of contact difference of potential** to thermionic and photo-electric emission are of a simple but most fundamental nature. These appear in the theoretical equation[95]

$$\phi_1 - \phi_2 = V_2 - V_1 + P_{12} \qquad (13)$$

in which $V_2 - V_1$ is the contact difference of potential between metals 1 and 2, and P_{12} is the Peltier coefficient at their junction. P is negligible in comparison with the other terms. This expression has been verified within the limits of accuracy of experiments for ϕ measured photo-electrically,[96] thermionic-

[90] Richardson, Phil. Mag. **31**, 149 (1916).

[91] Ladenburg, Ann. d. Physik **12**, 558 (1903); Partzch and Hallwachs, Ber. d. D. Phys. Ges. p. 749 (1907); Robinson, Phil. Mag. **32**, 421 (1916); Compton and Ross, Phys. Rev. **31** 374 (1919).

[92] Pohl, and Pringsheim, Verh. d. D. Phys. Ges. **12**, 349 and 682 (1910); **13**, 474 (1911); **14**, 46 (1912); Ives and Johnsrud, Astrophys. Jour. **60**, 231 (1924).

[93] Millikan and Souder, Proc. Nat. Acad. Sc. **2**, 19 (1916); Hughes, Bull. Nat. Res. Council Vol. 2, Part 2, No. 10, p. 115; Ives and Briggs, Bull. Am. Phys. Soc. **5**, 18 (1930).

[94] Frehafer, Phys. Rev. **15**, 110 (1920).

[95] Richardson, Phil. Mag. **23**, 261 and 615 (1912).

[96] Richardson and Compton, Phil. Mag. **24**, 575 (1912); Millikan, Phys. Rev. **18**, 236 (1921).

ally,[97] and as a heat of condensation of electrons.[98] Of these, the thermionic test is least accurate because it is difficult to make a contact difference of potential measurement between metals which are suitable for simultaneous determination of the thermionic constants.

In the case of metals with surfaces which are not atomically homogeneous, such as metals with surface patches of adsorbed materials, the simple relation above cannot hold, for the following reasons. The contact potential property of a composite surface, as measured by any of the standard methods, is simply an average value to which patches of different character contribute simply in proportion to their areas. The photo-electric work function, if measured as usual by the long wave-length limit, is characteristic of the most electropositive region on the surface. The thermionic work function, as measured by emission as a function of temperature, is again an average value, but an average in which the electropositive areas are weighted much more heavily

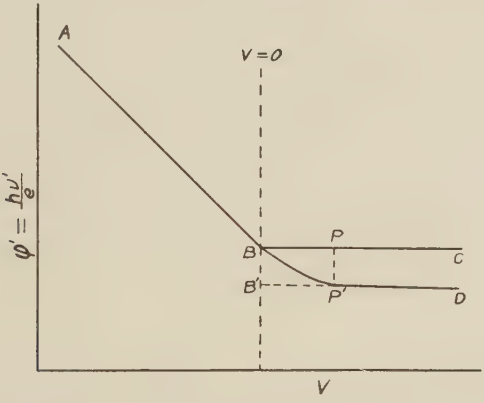


Fig. 1.5. Nottingham's curve of relation between "effective" low frequency threshold ν' and applied potential difference V. Curve ABC shows relation given by Einstein's equation and found for homogeneous metal surfaces. Curve ABD is found for activated surfaces. ϕ' is the "effective" work function.

than the electronegative areas. These differences are strikingly illustrated if one considers, as a simple example, a square centimeter surface of platinum on which is a patch of potassium of one square millimeter. The latter area determines the thermionic and photo-electric work functions, but affects the measured value of contact potential by only about one-half percent. Such differences have frequently been observed in dealing with activated emitting surfaces, but only the start of a systematic study of them has been made.[99]

Recent experiments and suggestions of Nottingham[99.5] appear to be very significant in this connection. He determined the maximum wave-length at

[97] Richardson and Robertson, Phil. Mag. **43,** 557 (1922).

[98] Van Voorhis, Phys. Rev. **30,** 318 (1928).

[99] Langmuir and Kingdon, Phys. Rev. **34,** 129 (1929).

[99.5] Nottingham, article soon to appear in Jour. Frank. Inst.; also Proc. Am. Phys. Soc. February (1930) meeting.

which photo-electric emission could be detected for each of a number of applied fields (both retarding and accelerating) between an emitting and a receiving electrode. The emitting electrode was of nickel, activated by a distilled film of an alkali metal. By Einstein's equation (12) the expected relationship would be as indicated by curve AB, Fig. 1.5. Here B is the point at which the field changes from retarding to accelerating and BC is the region of saturation current in which the threshold frequency is independent of field (except for the very slight Schottky effect described in the following section (B 5)). ν' is the lowest frequency at which photo-electric emission can be detected with applied potential difference V. The part AB of the curve has a slope of 45° if $h\nu'/e$ and V are expressed in the same units. Curve ABC actually represents the relationship as found with metals whose surfaces are homogeneous.

With activated surfaces, however, curve ABD describes the relations. The threshold frequency is changed by an accelerating field, PP' representing the reduction in "effective" work function due to an accelerating field BP. To illustrate the order of magnitude of this effect, an example is the following: A field BP of 4 volts caused a reduction $PP'=2$ volts in "effective" work function, but a further increase of 700 volts in the field caused only a few tenths of a volt further reduction.

A similar phenomenon is evident in the thermionic emission, at very small accelerating fields, from thoriated tungsten, as reported by Reynolds[105] (see Fig. 3 of Phys. Rev. **35**, 164 (1930)). There is every reason for believing that it is a general characteristic of emission from activated surfaces in weak accelerating fields.

There are several important consequences of this fact: (1) The values of work function of activated surfaces (Table IV) all refer to the "effective" work function as determined for accelerating fields greater than P' (Fig. 1.5) and extrapolated back to zero field (such as B'). Similarly photo-electric long wave-length limits are generally measured with considerable accelerating fields (see, for example, the work of Millikan). On the other hand, the long wave length limits calculated by Einstein's equation or extrapolated from the region AB of Fig. 1.5 give values characteristic of zero field, such as B. (2) There is a theory of change in effective work function produced by an accelerating field which accounts satisfactorily for experiments on emission from homogeneous surfaces. This theory, and all extensions of it which have thus far been proposed, fail signally in accounting for all the results obtained with activated surfaces. These theories are discussed in the following two sections (B 5, 6).

(5) **Electron emission in an accelerating electric field** never, strictly speaking, becomes completely saturated, for there is always an increase of current with field. For many purposes this effect is negligible, but in some cases it is very important. It gives rise to the necessity of examining just what we mean by the saturation electron emission, as defined for example by Eq. (6). It certainly plays an important role in phenomena at the cathode in some forms of gas discharges. It was first discussed by Lenard,[100] but was put into concrete

[100] Lenard, Ann. d. Physik **8**, 149 (1902).

and useful form by Schottky,[101] and is generally known as the *Schottky Effect*. Essentially the phenomenon is this:

Any escaping electron $(-e)$ is attracted back toward the metal by a force $F(x)$ due to the induced positive charge on the metal surface, which is *approximately* equal to the attraction of the "mirror image" charge $(+e)$. This force $F(x)$ is accurately $e^2/4x^2$ if the electron is at a distance x which is large compared with distances between consecutive atoms in the surface of the metal, for then the metal surface is effectively a conducting plane. Even

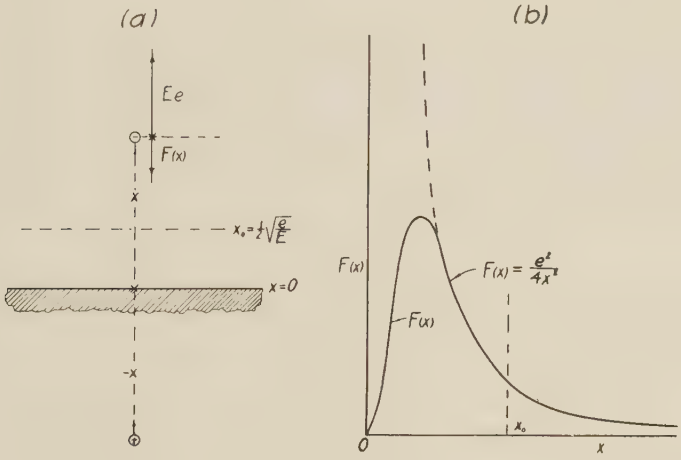


Fig. 2

$F(x)$ is the surface force which attracts an electron, distant x from the metal surface, toward the metal. x_0 is the distance at which this surface force is just equal and opposite to the force of the applied field.

at a distance of only four or five inter-atomic distances, this expression would hold to a rough approximation. On the other hand it would obviously fail completely at distances of the order of atomic diameters, being replaced, at the close distances, by a force function which depends on the discrete atomic structure of the metal surface and which must approach zero at the surface. The work function is given in terms of this force by

$$e\phi = \int_0^\infty F(x)dx. \tag{14}$$

Obviously $F(x)$ cannot be simply the image force $e^2/4x^2$ in the entire range from 0 to ∞, for in that case this integral would be infinite. If we make the natural assumption that $F(x)=0$ at $x=0$ and, for small displacements, increases linearly with x, then the simplest graphical representation of $F(x)$ is a parabola rising from $F(x)=0$ at the origin $x=0$ to a maximum and then down-

[101] Schottky, Phys. Zeits. **15,** 872 (1914); **20,** 220 (1919); Ann. d. Physik **44,** 1011 (1914); Zeits. f. Physik **14,** 63 (1923).

ward merging into the curve of the image force at $x = x'$, as shown in Fig. 2 b. With such a law of force Langmuir[102] has shown that Eq. (14) leads to

$$\phi = \frac{e}{2x'}$$

and that x' is of the order of an atomic radius. Values of x' in Angstrom units are: W 1.58, Pt 1.12, Mo 1.62, Ta 1.75, Th 2.13, Cs 3.95. These values vary roughly as the cube roots of the atomic volumes and are approximately equal to the atomic radii.

Direct experimental evidence regarding the nature of $F(x)$ is obtained by investigating the effect of accelerating fields on the electron emission. In the presence of an accelerating field E, there is a critical distance x_0 from the surface at which the resultant force vanishes. Within it the intrinsic surface attractive force $F(x)$ predominates, while beyond it this force is inferior to the applied force Ee. The critical distance is defined by the relation

$$F(x)_{x=x_0} = Ee. \tag{15}$$

For all except the strongest attainable applied fields it turns out that this critical distance x_0 is sufficiently large that the approximation $F(x) = e^2/4x^2$ is valid, so that we can use $e^2/4x_0^2 = Ee$, giving

$$x_0 = \tfrac{1}{2}(e/E)^{1/2}. \tag{16}$$

The range in which this approximation is justifiable may be suggested by taking an extreme case of $E = 1{,}000{,}000$ volts/cm. Then $x_0 = 19\ (10)^{-8}$ cm. The closest distances between atoms in a metal are[103] Th. 3.54 $(10)^{-8}$ cm W 2.73 $(10)^{-8}$ cm, K 4.50 $(10)^{-8}$ cm, etc. Thus even in this extreme case x_0 is about six times the inter-atomic distance. For $E = 10{,}000$ volts/cm the approximation would be very close indeed, for x_0 would be about sixty times the inter-atomic distance.

The applied accelerating field reduces the work which must be done by an electron in order to escape in two ways: (1) the integral of Eq. (14) needs to be carried out only to a distance x_0 instead of to infinity since, once past x_0, its escape is assured; and (2) the force which must be overcome in the path from 0 to x_0 is not $F(x)$, but $F(x) - Ee$. Calling ϕ' the "effective" work function, *i.e.*, the actual work which must be done by an electron to get free from the metal, we have

$$e\phi' = \int_0^{x_0} (F(x) - Ee)dx = \int_0^{\infty} F(x)dx - \int_{x_0}^{\infty} F(x)dx - \int_0^{x_0} Eedx.$$

The first integral is simply $e\phi$, and in the second integral $F(x) = e^2/4x^2$ in the range x_0 to ∞. Thus, introducing the value of x_0 from Eq. (16) we have

$$\phi' = \phi - (Ee)^{1/2}. \tag{17}$$

[102] Langmuir, Trans. Am. Electrochem. Soc. **29**, 157 (1916).
[103] Compton, "X-rays" p. 116 (Van Nostrand 1926).

The actual emission, whether photo-electric or thermionic, is governed by the "effective" work function ϕ' rather than the true work function ϕ. Thermionic emission lends itself most readily to a test of this Schottky effect, because of the relative ease of securing a large field at the surface of a filament of small diameter. Eq. (6) then takes the form

$$j = AT^2\epsilon^{-(e/kT)[\phi_0-(Ee)^{1/2}]} = AT^2\epsilon^{-e\phi_0/kT}\cdot\epsilon^{e(Ee)^{1/2}/T} = j_s\epsilon^{e^{3/2}E^{1/2}/kT} \tag{18}$$

or, if E is expressed in volts/cm,

$$j = j_s\epsilon^{4.389(E^{1/2}/T)}. \tag{19}$$

This equation may easily be tested if the current is small and gas ionization absent (so as to avoid distortion of the field by space charge, *see* Part II) for then E is directly proportional to the applied difference of potential V, the constant of proportionality being calculable by principles of electrostatics and depending only on the geometry of the electrodes. Grouping together this constant and the quantity $e^{3/2}/kT$ as a "shape factor" S, we have

$$j = j_s\epsilon^{SV^{1/2}}, \quad \text{whence} \quad \log j - \log j_s = SV^{1/2}. \tag{20}$$

The graph of $\log j$ against $V^{1/2}$ should therefore give a straight line, whose slope S is calculable from the geometry of the electrodes and whose intercept $\log j_s$ at $V=0$ serves to define the "saturation" current J_s.

Experimental tests of Eq. (20) are very illuminating in that they afford an insight into surface conditions which greatly affect electron emission. With clean metal filaments Eq. (20) is accurately verified in the entire range from 2 or 3 volts (below which initial velocities of emission and current limitation by space charge may cause a departure) up to the highest fields (10^6 volts/cm) which have been used in thermionic emission.[104] Generally the shape factor S is from 5 percent to 100 percent larger than the value calculated from the geometry, but this departure is found to be due to microscopic roughness of the surface which concentrates the field at protruding points. Abnormally large values of the slope S are found for wires which have been subjected to positive ion bombardment, indicating a "pitting" of the surface when struck by high speed ions. With clean tungsten wire of small crystal grain structure and flashed at a high temperature to smooth out surface inhomogeneities Eq. (20) is satisfied as accurately as the measurements can be made, with the theoretical value of the shape factor S. This proves that the surface force $F(x)$ is identical with the electrostatic image force $e^2/4x^2$ at distances greater than the minimum value of x_0 attained in these experiments.

Even if the surface force should, for any reason, not be the image force, the above method may still be applied to explore the surface force if it is permissible to assume that the force is uniform over a plane parallel to the surface. Writing Eq. (6) in the most general form suggested by Eqs. (14) and (17).

[104] Schottky, Ann. d. Physik **44**, 1011 (1914); Dushman et al, Phys. Rev. **25**, 338 (1925); Lauriston and Mackeown, *ibid.* **32**,326 (1928); Phorte, Zeits.f. Physik **49**, 46 (1928); de Bruyne, Proc. Roy. Soc. **A120**, 428 (1928); Reynolds, Phys. Rev. **35**, 158 (1930).

$$j = AT^2 \exp\left\{-\frac{1}{kT}\int_0^{x_0}[F(x)-Ee]dx\right\}$$

and differentiating by E, we obtain

$$\frac{d\log j}{dE} = \frac{ex_0}{kT}. \tag{21}$$

Thus the rate of change of the logarithm of the thermionic current with change in accelerating field gives the distance x_0 from the surface at which the surface force $F(x)$ equals the applied force. By using a wide range of fields, the surface force is thus found over a wide range of distances from the surface.

In general the emission from activated surfaces, such as caesium or thorium on tungsten or Wehnelt oxides, does not follow the simple Schottky equation (19), but shows wide departures especially at low fields. Furthermore these departures are greater for surfaces only partly covered with activating material than for surfaces which are either nearly bare or nearly completely covered. Eq. (21) has been used to investigate the peculiar surface force for such surfaces, as will be described later. It may be noted, further, that the degree of approximation to the Schottky Eq. (19) is largely affected by the previous history of the surface as regards method of activation, ion bombardment, etc.[105]

The only present theory of these departures from Schottky's equation is based on a suggestion by Langmuir[106] that the electropositive atoms on the activated surface give rise to electric force fields near the surface which vary from point to point and affect differently the escape of electrons from different regions, thus tending to produce imperfect saturation. The electropositive atoms thus behave as a sort of positively charged "grid" to affect the emission. These electropositive atoms may act in groups, or "patches" which are electropositive with respect to the surrounding bare metal, in which form the theory has been utilized by Richardson and Young[107] and by Reynolds.[105] Or they may act individually as charges situated just outside the surface giving rise, with their mirror images, to doublet fields, in the form of the theory developed by Becker.[108] We propose to discuss these theories somewhat at length, both because of the interest attaching to the phenomena and because the following analysis shows that *none* of the present theories appears to be capable of accounting for the observed departures from Schottky's equation.

The theory of patches was suggested by Langmuir[106] to explain the lack of definite saturation observed with filaments whose surfaces are only partly covered with adsorbed films, such as those of thorium. At 1500°K the electron emission from a tungsten surface completely covered with a layer of thorium atoms is about 126,000 times as great as that from pure tungsten.

[105] Reynolds, Phys. Rev. **35,** 158 (1930).

[106] Langmuir, Gen. Elec. Rev. **23,** 504 (1920).

[107] Richardson and Young, Proc. Roy. Soc. **107,** 377 (1925).

[108] Becker and Mueller, Phys. Rev. **31,** 431 (1928); Becker, Bell Tech. Lab. Reprint B-412, August (1929).

If half of a large tungsten surface were covered with such a thorium film while the other half were bare, the total emission would be about 63,000 times that of pure tungsten. If the thorium, however, is uniformly distributed over the filament, the work function is a linear function of the amount of thorium on the surface, and thus, according to Eq. (7.5), the electron emission from the surface would be 356 times that of pure tungsten $(126{,}000)^{1/2}$.

In the first case we have an emission which corresponds to an arithmetical average while in the second we have to deal with a geometrical mean. In either case the current should increase with the applied accelerating field in accord with Schottky's theory, Eq. (20). Thus with enough thorium to cover half the surface the line AB in Fig. 3(a) represents the emisson as a function

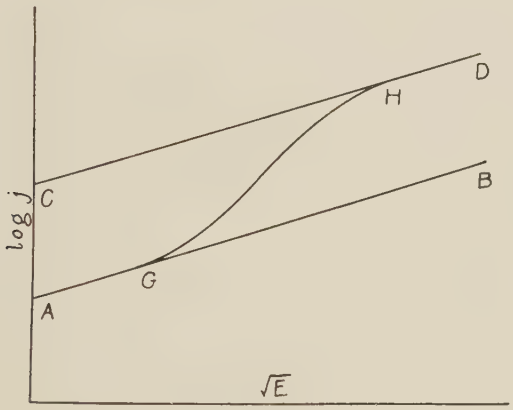


Fig. 3. Effect of patch size on electron emission.

of field for uniform distribution while a parallel line CD represents the 356 times larger currents when the thorium is gathered into large patches. If the linear dimensions b of a patch of active material on a cathode are sufficiently small compared to x_0, the distance to the critical surface as given by Eq. (16), the distribution will be effectively uniform and AB will give the current, but when $b \gg x_0$, CD will apply. As the field increases, x_0 decreases and thus the transition from AB to CD may be expected to occur as illustrated by $AGHD$ so that in region GH the Schottky effect will be abnormally large.

A quantitative estimate of the field strength that is needed to cause the emission to depart appreciably from the Schottky line AB, for example at G, may be made by considering the following simplified case. Let a surface consist of patches of two kinds having a contact potential difference V_0 between them, and let them be arranged in alternate squares like a checkerboard, each square having sides of length b. If the potential of the more negative squares is 0, and there is no applied field, the distribution of potential in the space near the checkerboard is given by the following expression

$$V = \frac{1}{2}V_0 + \frac{4V_0}{\pi^2} \sum_{n=1,3,5} \left[\frac{1}{n^2} \exp\left(-n(2)^{1/2}\pi z/b\right) \cos\frac{n\pi x}{b} \cos\frac{n\pi y}{b} \right]$$

where z is the distance from the surface while x and y are the distances measured from the center of one of the more electropositive squares in directions parallel to the edges of the squares.

To determine the motions of electrons we must add together this potential V, the potential due to the external field E, and the fictitious potential $e/4z$ due to the electric image force. The name "motive" has been proposed[109] for a scalar quantity whose gradient in any direction and at any point represents the force component per unit charge which *acts on an electron or ion.* The motive M is thus

$$M = M_\infty + \frac{e}{4z} + Ez + V$$

where M_∞ is a negative quantity numerically equal to the work function for 0 field and expressed in volts. Thus, for tungsten $M_\infty = -4.52$ volts.

By differentiating M with respect to z and equating to 0, an equation is obtained which, when solved for z, gives the distance z_M at which the motive is minimum; that is, it gives the location of the critical surface. This distance z_M is a function of x and y, but if V_0 is placed equal to 0, z_M then has the same value as x_0 given by Eq. (16). The value of z_M corresponding to $V_0 = 0$ we will call z_0 and we may then place

$$z_M = (1+\lambda) z_0 \tag{22}$$

where λ is a small quantity if $z_0 \gg b$. The equation $dM/dz = 0$ can then be expressed in the folowing form,

$$1-(1+\lambda)^{-2} = \alpha \exp(-\beta\lambda)$$

where

$$\alpha = \alpha_0 \cos\left(\frac{\pi x}{b}\right) \cos\left(\frac{\pi y}{b}\right) \tag{23}$$

$$\alpha_0 = \frac{4\,(2)^{1/2} V_0}{\pi E b} \exp(-\beta) \tag{24}$$

$$\beta = \frac{\pi}{b}\left(\frac{e}{2E}\right)^{1/2} = \frac{4.44 z_0}{b}. \tag{25}$$

This expression can be expanded into a power series in λ, and then, by reversion, the following series is obtained for λ in terms of α,

$$\lambda = \frac{1}{2}\alpha + \frac{\alpha^2}{8}(3-2\beta) + \frac{\alpha^3}{16}(5-9\beta+3\beta^2).$$

This equation thus allows us to determine by Eq. (22) the distance z_M to the critical surface of minimum motive as a function of x, y and V_0. The motive M_M at the critical surface can also be expressed in terms of λ and therefore in

[109] I. Langmuir and K. H. Kingdon, Proc. Roy. Soc. **A107**, 68 (1925).

terms of α. The final expression for the change in motive at the critical surface produced by making E different from 0 is given by,

$$\Delta M_M = (eE)^{1/2}\left[1+\frac{\alpha}{2\beta}-\frac{\alpha^2}{8}-\frac{\alpha^3}{16}(1-\beta)\right].$$

The total electron emission may then be calculated* from Eq. (6) by decreasing the value of b by an amount equal to $(e/kT)\Delta M$ and averaging the resulting current density over the whole surface of the cathode. The terms involving odd powers of α then drop out and the total current density from the surface is found to be,

$$j = j_n\left[1+\frac{\alpha_0^2}{32}\left(\frac{1}{\beta^2}-1\right)\right] \tag{26}$$

where j_n is the current density corresponding to the normal Schottky effect as given, for example by Eq. (19).

For convenience, the value of α_0 may be calculated by eliminating E between Eqs. (24, 25). Thus

$$\alpha_0 = 2.55\times 10^6 V_0 b\beta^2 \exp(-\beta) \tag{27}$$

where V_0 is expressed in volts.

As an example to illustrate the use of these equations we may take the case of a tungsten filament having a radius 0.01 cm inside of a cylinder having a radius of 1 cm.

With the cylinder at various positive voltages V_a with respect to the filament as given in line 1 of Table VI the accelerating field strengths E at the filament are given in line 2. Line 3 shows the normal increase in electron emission according to the Schottky theory as given by Eq. (19). The normal distance z_0 of the critical surface from the cathode is in line 4.

TABLE VI. *The effect of local fields due to checkerboard patches of thorium-covered tungsten in causing lack of saturation. Size of squares $b=10^{-6}$ cm; contact pot. diff. $V_0=1.9$ volts.*

1 V_a (volts)	0	100	250	500
2 E (volts cm^{-1})	0	2170	5430	10850
3 j_n/j_s	1.000	1.146	1.242	1.357
4 z_0 (10^{-6} cm)	∞	4.0	2.6	1.8
5 β	∞	18.0	11.4	8.0
6 α_0	0	2.10^{-5}	0.007	0.10
7 $(j-j_n)/j_n$	0	-10^{-10}	-10^{-6}	-3.10^{-4}
8 λ_M	0	10^{-5}	0.003	0.05

Let us now assume that the surface of the filament is one-half covered in checkerboard fashion by square patches of close-packed thorium atoms, the side of each square being $b=10^{-6}$ cm. Since a tungsten surface completely

* This assumes that the electrons which reach the critical surface from the cathode have a Maxwellian velocity distribution. This will be true regardless of the motive distribution near the cathode, if the critical surface is approximately plane, that is, if $\lambda \ll 1$. See Mott-Smith and Langmuir, Phys. Rev. **28**, 754–8 (1926) particularly the theorem on p. 756.

covered by thorium contains[110] about 7×10^{14} atoms of thorium per cm^2, each patch contains 700 atoms.

The values of β and α_0 given by Eqs. (25) and (27) are in lines 5 and 6 while line 7 gives the fractional change in current caused by the local fields of the patches as calculated by Eq. (26).

Thus even with such extremely non-uniform distribution of thorium atoms as we have assumed, with clusters of 700 atoms, the local fields cannot cause variations of emission which are at all comparable with those of the normal Schottky effect.

With patches of a little larger size, containing about 7000 atoms and about 3×10^{-6} cm square, the values of $(j-j_n)/j_n$ calculated by Eq. (26) become much larger and reverse in sign, so that the Schottky effect will increase rapidly as illustrated by GH in Fig. 3 but then $1/\beta$ and $\beta\lambda$ become so large that the approximations made in deriving these equations are no longer valid. However, it is very improbable that such large clusters of adsorbed atoms can exist on plane surfaces. Becker has shown,[108] for example, that a thorium layer at emission temperatures behaves like a 2-dimensional gas on the surface.

A random distribution of adsorbed atoms would also give considerable irregularities in emission even over rather large areas. For example, if the average number of atoms in equal square patches (each of area b^2) is q, then the average difference between the number of atoms in any two squares selected at random is $(2q)^{1/2}$ so that the average contact difference of potential between adjacent squares is $5.2\times10^{-8}\ V_0/b$ where V_0 corresponds to the difference between completely covered and uncovered surfaces. Thus if we take $b=4\times10^{-6}$ (about equal to z_0) and $V_0=1.9$ we get a potential difference of only 0.025 volt. With a filament at 1500° this would cause a 20 percent difference in the normal emissions of the two surfaces. In this case the arithmetic mean and the geometric mean (of 0.9 and 1.1) differ by only 0.5 percent so that no abnormally large Schottky effect can occur.

Finally, if there are attractions between the thorium atoms so that they form clusters of n atoms each, the clusters being distributed at random, then the average contact difference between adjacent squares of areas b^2 becomes $\Delta V=5.2\times10^{-8}\ V_0\ n^{1/2}/b$. Thus the difference between the arithmetic and geometric means which is proportional to $(\Delta V)^2$ becomes 50 instead of 0.5 percent if $n=100$.

Langmuir, Richardson and Becker have experimentally observed cases with filaments whose surfaces are partly covered by adsorbed films where the variation of emission with voltage is much greater than the variation observed with bare surfaces or completely covered surfaces.

K. H. Kingdon and Langmuir in some unpublished work in 1923 and more recently Reynolds[105] measured the emission as a function of voltage and temperature for thoriated filaments of various activities. They had unusually good vacuum conditions, the metal parts being thoroughly degassed and the

[110] I. Langmuir, Phys. Rev. **22,** 375 (1923).

bulbs excessively baked out and immersed in liquid air. The filaments were in the axis of a cylinder provided with end guards so that the field strengths were accurately known.

Reynolds, working with thoriated filaments which had been flashed to very high temperature and then activated, found a variation with voltage in excellent agreement with the Schottky theory for field strengths exceeding about 10,000 volts/cm, as shown by curves A and B of Fig. 4 for two different degrees of activation. If, however, the filament had been well activated and was then slightly deactivated by positive ion bombardment (which occurred slowly if the liquid air were removed while high voltage was applied), then the emission did not agree with the Schottky curve above 10,000 volts/cm, as

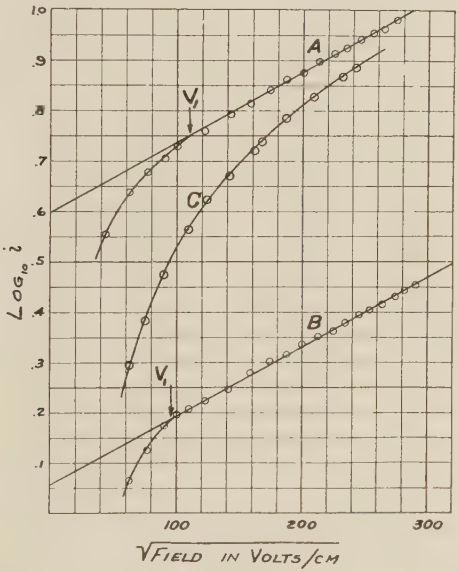


Fig. 4. Electron emission from a heated thoriated filament as a function of field, showing approach to Schottky's curve A, B above 10,000 volts/cm for a regularly activated filament, and the different relation C for a filament which had been subjected to positive ion bombardment. (Reynolds)

shown by curve C, Fig. 4. This bombardment must have roughened the surface and there is evidence that it also fractured the surface layer of tungsten crystals. Subsequent flashing at high temperature followed by activation always brought the filament back to the condition giving curves A and B.

Kingdon's and Langmuir's results were the following. The thoriated filament was flashed 1 minute at 2740° and activated for 25 minutes at 2120°. It was free from carbon and thus could be completely deactivated by heating 10 seconds at 2600°K; partial deactivation was accomplished in 4 stages by heating 30 or 60 seconds at 2300 or 2400°. In each stage the emission was measured at 5 voltages and about 6 testing temperatures. Fig. 5 gives a few of the typical curves obtained by plotting $T \log_{10} j$ against $\sqrt{E}$ where E is in volts per cm. According to Eq. (19) the lines obtained should be straight and

should all have a slope 1.906. At the highest field strengths (about 10,000 volts/cm) the curves are seen to approach the theoretical slope. In Fig. 5 they have been displaced in a vertical direction by various amounts Δ in order to show how they approach the Schottky line asymptotically.

These results indicate that with nearly complete thoriation of the surface ($\theta = 0.91$) and with a bare surface ($\theta = 0.00$) the approach to the Schottky curve is fairly close,* but relatively large departures occur with incomplete thoriation ($\theta = 0.69$). In every case the departures from the Schottky line

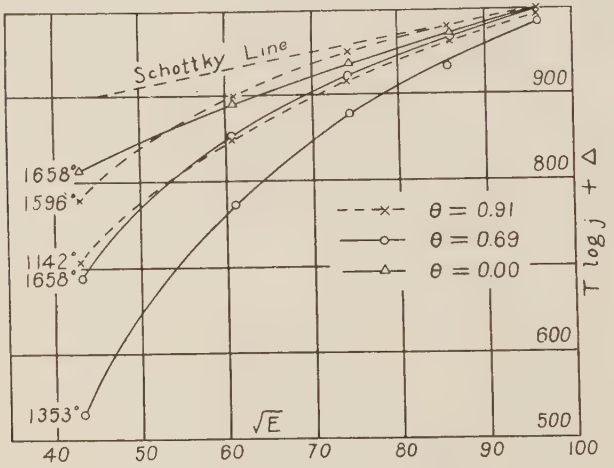


Fig. 5. Electron emission from thoriated filament as a function of activity θ and field, at various temperatures. (Kingdon and Langmuir)

become greater as the temperature is lowered, thus proving that the deviatons are not due to space charge.

Table VII gives η, the ratio of the observed emission at a field of 1840 volts/cm (50 volts on anode) to that calculated from the Schottky line by extrapolating down from $E = 9200$ as shown in Fig. 5. It is seen that the maximum departure from the Schottky law occurs when the tungsten is about 70 per cent covered by thorium.

TABLE VII. *Departures (η) from Schottky's law with field of 1840 volts/cm.*

θ	$T = 1500°$ η	j(amps cm^{-2})	η at a temp. giving $j = 17 \times 10^{-6}$ amps cm^{-2}
0.91	0.81	9200×10^{-6}	0.69
0.76	0.64	1800 "	0.46
0.69	0.59	765 "	0.44
0.24	0.72	3.75 "	0.77
0.00	0.80	0.23 "	0.89

* Closer fits than this to the Schottky curve and over a much larger range of fields have been obtained by Reynolds with bare wires. The small departure here is believed to be due to the effect of slight surface roughness on the "shape factor" of Eq. (20).

When tungsten filments are heated to very high temperature so that evaporation occurs, the etching of the surface causes the development of the dodecahedral faces. These form angles of 120° with each other. Sometimes the resulting ridges are large enough to see under a microscope. The effect of such ridges and valleys on the variation of emission with voltage can be roughly calculated from the theory that we have developed for the case of patches. If we consider a plane surface level with tops of the ridges, then the potential at points on the surface over the centers of the valleys will be roughly Fh where h is the effective height of the ridges above the valleys. We may therefore use Eqs. (26, 27) by merely substituting $V_0 = Fh$. The increase of V_0 with F will then have the effect of causing departures from the normal Schottky effect which increase still more rapidly with increasing voltage than in the case of patches in a plane surface. The effect of such geometrical irregularities in the surface would thus seem also to give a curve of the type *AGHD*, Fig. 3.

The final result of this study is the conclusion that the "patch" theory predicts variations of emission with field which, while qualitatively in the right direction, are found on closer examination to be totally unlike the observed variations, such as shown in Figs. 4, 5. There are two outstanding difficulties. In the first place, to obtain departures from the Schottky curve comparable with those observed, the patches must be assumed to contain many thousands of atoms. In the second place, the patch theory predicts a departure from the Schottky curve which is small with small field and increases with large fields, whereas exactly the reverse is the actual case.

The reason for this failure is easy to see. By Eq. (21) it is seen that the rate of increase of current with field depends solely on the conditions at the critical x_0 (or later z_0) where the intrinsic surface force is just balanced by the applied field. For weak fields x_0 is large, approaching ∞ for zero field. Obviously local inhomogeneities in the field due to "patches" or to adsorbed ions, which may be important close to the surface, average out and become negligible at larger distances. Thus any theory based upon such surface inhomogeneities must fail to account for the observed relatively large deviations from the Schottky equation at small fields, and would predict large deviations at strong fields, which again is contrary to the facts. In order to obtain appreciable departures from the Schottky equation, these theories require that the scale of the surface inhomogeneity be comparable with the critical distance x_0, hence that the patches be large enough to contain thousands of atoms. Even if this were true, which is very unlikely, we would still be faced with the fundamental failure to account for the fact that it is at low rather than high fields that the deviation from the Schottky equation is most pronounced in this case of activated surfaces.

We now turn to that form of "grid" theory of activated surfaces which has been interestingly investigated and extended by Becker,[108] and which we may call "*the adion field theory*," borrowing from Becker the term "adion" to designate "adsorbed ion." The theory aims to account for the three peculiarities of

activated surfaces, (1) low work function, (2) lack of saturation in emission, (3) electropositive character of surface, as follows:

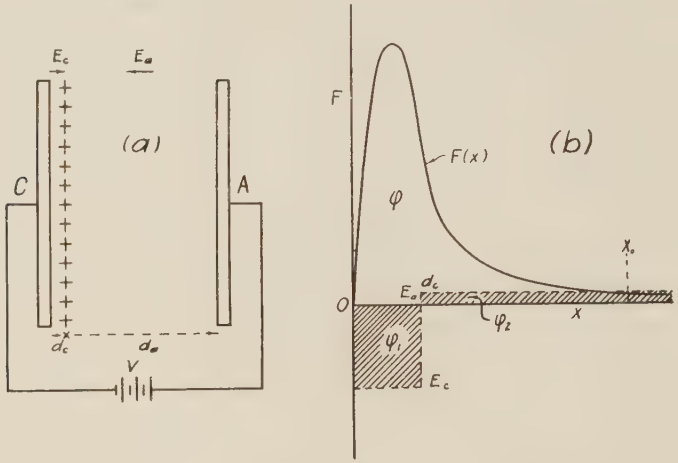


Fig. 6. Adion theory of activated surfaces.

Consider two similar electrodes C and A (Fig. 6(a)) distant d apart, and assume the cathode C to be more or less covered with a surface layer of positive ions distant d_c. We may suppose the electrodes to be of tungsten and the ions to be of some activating material such as Ba^+. The fact that the barium evaporates as ions is evidence of its existence in the form of ions on the surface, since the surface forces which tend to ionize it are presumably strongest at or very near the surface. The effect of this layer of "adions" on the field between the plates may be considered to *a first approximation* as that of a layer of positive charge of surface density $\sigma = Ne$, where N is the average number of adions per cm². The field due to this layer is uniform and is very nearly $E_c = 4\pi Ne$ on the cathode side and $4\pi Ne\, d_c/d_a$ on the anode side. The potential rise from cathode to layer is $4\pi Ne\, d_c$, and is equal to the potential drop from layer to anode.

If the electrodes C and A were connected by a wire ($V = 0$) then the adion layer would be without effect on the emission from the cathode C, since the total work done on an electron in escaping from C to A is the same as if the adion layer were absent. However the field E_a is interpreted by an observer as a contact difference of potential between C and A and he compensates it by an equal opposite field V applied externally in order to obtain what he considers to be the true saturation emission from C with zero field. (This is exactly in accordance with observation and procedure in any case of emission beween dissimilar electrodes.) Under these conditions the work which an electron must do to escape is less, by this apparent contact difference of potential $V = E_a d_a = E_c d_c$, than in the absence of the adion layer. The emission is therefore greater in accordance with this decrease in effective work function.

These considerations are shown graphically in Fig. 6(b), where $F(x)$ represents the surface force at various distances x from the cathode and the area under it equals the work function ϕ of the normal cathode surface. The adion layer produces an additional force E_c inside the distance d_c, and E_a in the reverse direction beyond this distance; these forces add an amount ϕ_1 to the energy of the escaping electron within the distance d_c and subtract an equal amount ϕ_2 beyond. The total work done in escaping is thus unchanged, being $\phi-\phi_1+\phi_2=\phi$. If however, the outer field E_a is compensated by applying an external field as in correcting for contact potential difference, the effective work function is reduced to the difference between the areas $\phi^1=\phi-\phi_1$. This theory therefore gives a plausible interpretation of the effect of adions on the work function and the contact difference of potential.

Thus far the theory fails to account for departures of emission with accelerating field from Schottky's Eq. (19). In an attempt to do this, we go to a second approximation and treat the adion layer no longer as a uniform positive layer, but as discrete positive charges, each of which with its negative mirror image in the metal constitutes an electric doublet. Electrons escaping at various points pass through fields of various intensities and extending into the region outside the adion layer. Thus the effective work function varies over the surface and the nature of the force F varies with x differently than in the cases of bare or of uniformly covered surfaces. Qualitatively this is what is needed to account for the departure from the Schottky equation. Quantitative investigation, however, shows that this theory fails just as did the "patch" theory in that it fails to account for the fact that the departure from Schottky's equation increases with decreasing fields. In Fig. 6(b), for example, such doublet fields may be important at distances of d_c or $2d_c$ from the surface, but not at the relatively large distances x_0 which, by Eq. (21) determine the variation of current with field for all ordinarily attainable fields.

Dr. Becker has kindly informed us of his intention soon to publish, in the Physical Review, an analysis of the "adion" theory extended to treat the case of an *activated rough* surface, rather than the ideal activated infinite plane discussed above, and has stated that this analysis will be found to be in much better agreement with the facts. The "rough" surface is conceived of as composed of the surface crystal facets, whose linear dimensions may be of the order of the critical distance x_0 for weak fields.

In connection with this theory there is the further interesting question as to what extent the specific emission properties of the adsorbed material are present in a monatomic layer. A layer ten, or five, atoms deep of caesium on tungsten almost certainly possesses the emission properties of caesium. To what extent do these specific properties completely vanish for a monatomic layer, being replaced only by an electrostatic effect of the charges of the ions?

In conclusion, it may be well to point out one assumption underlying both these forms of "grid" theory, *i.e.*, that the explanation of the departure from Schottky's theory in the case of activated substances is to be sought in the influence of adsorbed layers on the work function, neglecting a possible peculiarity in the constant A of the thermionic Eq. (6). For example Eq. (21)

which has been used by Becker and Mueller[108] to calculate the nature of the surface force $F(x)$ is based upon the assumption that A is constant. However Fowler[111] has recently shown that A is not, in general, a constant, and that particularly in the case of activated surfaces there may be a marked dependence of A upon temperature. We shall see in the next section that a strong electric field may affect electron emission (in addition to thermionic emission) in the same way in which high temperature affects thermionic emission. It may be, therefore, that activated surfaces have such a dependence of A upon the field as to account for the abnormally small emission in weak fields.

(6) **Field currents.** According to experimental evidence, the Schottky theory, Eq. (18), becomes more accurate with high field strengths. It is evident from Eq. (18) that if the field should increase until

$$Ee=\phi_0^2 (\text{giving } E_M=\phi_0^2/e) \tag{28}$$

the work function and cooling effect should vanish and the current density should increase to a limiting value,

$$j(\text{max.})=AT^2. \tag{29}$$

We may measure rates of increase of the current j for given relative increases in T or E by exponents n_T and n_E defined as follows:

$$\text{and} \qquad \begin{aligned} n_T &= d(\log j)/d(\log T) \\ n_E &= d(\log j)/d(\log E). \end{aligned} \tag{30}$$

By logarithmic differentiation by Eq. (18) with respect to T or E we then obtain

$$\begin{aligned} n_T &= 2+(e/kT)\left[\phi_0-(Ee)^{1/2}\right] \\ &= 2+(b-4.389E^{1/2})/T \end{aligned} \tag{31}$$

and

$$n_E=\tfrac{1}{2}(e/kT)(Ee)^{1/2}=2.20E^{1/2}/T. \tag{32}$$

For tungsten $b=52{,}600$. At $T=1500$, according to Eq. (31), n_T decreases from 37 when $E=0$ to $n_T=2$ when E approaches the limiting value, E_M, given by Eq. (28) which may for convenience be expressed

$$E_M=0.0519b^2. \tag{33}$$

For tungsten the limiting field is thus

$$E_M=1.44(10)^8 \text{ volts cm}^{-1}. \tag{34}$$

At this field the current density, by Eq. (29) with $A=60.2$, would be $1.36\ (10)^8$ amps. cm^{-2}.

According to Eq. (18) it should be possible even at room temperature ($T=300°$) to get a current of $5.4\ (10)^7$ amp cm^{-2} by raising the field strength to the same maximum value E_M.

[111] Fowler, Proc. Roy. Soc. **A122,** 36 (1929).

The temperature coefficient of the electron currents that could thus be pulled out of metals by intense fields becomes negligible compared to the temperature coefficients ordinarily characteristic of electron emission.[112]

Eq. (32) shows that the relative increase in current with field strength increases as the field becomes greater until finally, as E increases up to E_M, $n_E = b/2T$. Thus with tungsten at room temperature, the maximum value, n_E, is 88. As the field approaches the value E_M, a one percent increase in field causes a 2.4 fold increase in j. When E is 90 percent of its final value, that is, when $E = 0.9\ E_M$, the emission is only about 10^{-4} as great as when $E = E_M$.

If there are slight geometrical irregularities, such as ridges or crystal corners on the surface of the cathode, there will be much stronger fields at the summits of these elevations. In regions where the field is only 10 percent greater, the current density would be so much greater that practically the whole emission may well occur from such points.

With the strong fields necessary to pull electrons out of cold metals, the electrostatic pull exerted on the surface may become very great. This force amounts to $4.4\ (10)^{-7}\ E^2$ dynes cm^{-2} when E is in volts cm^{-1}. Thus, when $E = E_M$ for tungsten, the pull on the surface is equivalent to a negative pressure of about 9000 atomospheres. This, however, is a force far below the tensile strength of a material like tungsten and therefore should not be able to disrupt the surface unless the temperature is extremely high or there are projections only loosely held to the underlying metal.

Experimental phenomena which involve the pulling out of electrons from cold metals have been observed and studied in considerable detail in recent years. R. W. Wood[113] passed discharges from an induction coil between platinum spheres 1.5 millimeters in diameter placed 1 to 5 millimeters apart in relatively high vacuum. X-rays were produced at the anode sphere, showing that the voltage across the gap was very high. This is clearly an example of a cold cathode discharge. There was a considerable transfer of platinum from the anode to the cathode. This was probably caused by the mechanical disintegration of the anode resulting from the temperature fluctuations produced by the intermittent discharge, such as is produced at the anode of a Coolidge X-ray tube when operated on alternating current. The mechanical force due to the high local field would pull the disintegrated metal to the cathode. On reversing the polarity showers of minute sparks were shot out from the new anode, consisting of incandescent particles ripped off the anode.

F. Rother[114] has investigated the electron emission from a cathode in an intense field obtained by bringing the surfaces of the cathode and anode extremely close together, the distances ranging from 10^{-6} to 10^{-3} centimeter. In this way it is possible to obtain great field strengths with such low voltage differences that measurements can be conducted at atmospheric pressure.

[112] W. Schottky and H. Rothe, Handbuch d. Expl. Physik Vol. 13, Part 2, p. 261, Leipzig 1928.

[113] Wood, Phys. Rev. **5**, 1 (1897).

[114] Rother, Ann. d. Physik **81**, 317 (1926).

Lilienfeld, using high voltages and high vacuum, utilized electron currents drawn from pointed cathodes and produced X-ray tubes on this principle. Millikan and Eyring[115] have measured quantitatively the electron currents drawn from tungsten filaments with fields ranging from 0.4 to 1.1×10^6 volts per centimeter, and have studied the effect of the temperature of the filament and the condition of its surface.

The past history of the surface has a very great effect on the field strength needed to draw electron currents. A fresh surface generally shows large emission which may vary with time so that the results are at first not reproducible. If the surface is "conditioned" by drawing currents with a high field, then the emission at lower fields usually becomes reproducible. Glowing the filament to 2700°K changes the surface in such a way as greatly to increase the field required to obtain a given electron current with the filament at room temperature. These observers found not only that the temperature coefficient of the electron emission at high fields is much less than at low fields, but obtained the surprising result that the "field currents" were entirely independent of the filament temperature between 300°K and 900°K, within the experimental error of about 5 percent. At 1100°, however, the field needed to draw a given current (10^{-11} amperes) was as much as 30 percent less than the field needed at 300° when the condition of the filament was such that strong fields were needed. When, however, the condition was such as to give greater electron emission (fresh filament), the critical field even at 1100° was the same as at 300°.

In a later paper, Eyring, MacKeown, and Millikan[116] studied field currents in high vacuum from points of tungsten, platinum, nickel and steel. The points were made as nearly as possible in the form of hyperboloids of revolution and were brought close to a plane tungsten disk as anode. The field strength at the surface of the point was calculated according to a theory given. Within the experimental error (2 percent), the field at the surface required to draw a given current was independent of the distance between the point and plane, and thus was independent of the total voltage applied between these electrodes.

In all of these results as well as those published by the research staff of the General Electric Company, London,[117] the data agreed with the empirical equation[118]

$$j=A(T+cE)^2\epsilon^{-b/(T+cE)} \tag{35}$$

according to which for strong fields and low temperatures the logarithm of the electron current is a linear function of the reciprocal of the field strength. This type of relation is in much better agreement with the experiments than that suggested by Schottky's theory according to which log I is a linear function of $E^{1/2}$.

[115] Millikan and Eyring, Phys. Rev. **27**, 51 (1926). This article contains references to the earlier investigators in this field such as Earhart, Kinsley, Hobbs, Hoffman and Lilienfeld.

[116] Millikan and Eyring, Phys. Rev. **31**, 900 (1928).

[117] Research Staff of the Gen. Elec. Co., London, Phil. Mag. (7) **1**, 609 (1926).

[118] Millikan and Lauritsen, Proc. Nat. Acad. of Sci., **13**, 45 (1928).

The fact that the temperature coefficient of the field current is practically zero and is thus negligible as compared with that given by the classical theory, Eq. (29), is proof that the electron energies do not increase in proportion to the absolute temperature as they should do according to the classical gas laws. This fact, however, seems to be consistent with the modern quantum theory of the degenerated electron gas as developed by Fermi, Dirac, Nordheim, and Sommerfeld.[119] On this basis there has been developed a theory[120] according to which the field currents from cold metals should be given by

$$j = aE^2\epsilon^{-D/E} \tag{36}$$

where a and D are constant for the metal. The data of Eyring, MacKeown and Millikan[116] obtained with a steel point give straight lines, when $\log j$ is plotted against $1/E$, whose slope gives for D the value $D = 6.7\ (10)^6$ volts cm^{-1}. If we assume that the value of b for steel is 67000, this would give for the value of c in Eq. (35) the value 0.010 deg. cm $volt^{-1}$.

By logarithmic differentiation of Eq. (36) we find for the value of n_E, defined by Eq. (30), the value

$$n_E = 2 + D/E. \tag{37}$$

For the experiments with the steel points described just above, for which the average field strength was $E = 0.33\ (10)^6$, $n_E = 22$.

In some unpublished work by W. D. Coolidge and I. Langmuir in 1922 a V-shaped tungsten filament 0.0216 cm diameter was mounted in a small bulb opposite the center of a molybdenum disk 3 cm in diameter, the distance being adjustable by a micrometer through a Sylphon joint. The molybdenum disk was baked out at incandescence and the filament was aged at about 2700°K. One leg of the filament was then broken off by bringing the plate in contact with it so that there was then a straight piece of filament opposite the molybdenum disk.

The voltage difference between point and plane was adjusted until a field current of 10^{-5} amperes was obtained from the point at room temperature. The distance was varied and in each case the voltage was changed until the current rose to the same value. Fig. 7 gives the results obtained in this way, the voltages being plotted as abscissas and the logarithm of the distance between the end of the wire and the plane being given by the ordinates.

Maxwell[121] has calculated the potential distribution between two coaxial confocal paraboloids of revolution maintained at a given difference of potential. The potential at any point P is found to be a linear function of $\log (r^{1/2} \cos \theta/2)$ where r is the distance of the point P from the focus O and θ is the angle POT where the line OT lies in the axis of the paraboloid. The electric

[119] E. Fermi, Zeits. f. Physik **36,** 902 (1926); P. A. Dirac, Proc. Roy. Soc. London, **112,** 661 (1926); L. Nordheim, Zeits. f. Physik **46,** 833 (1928); A. Sommerfeld, Zeits. f. Physik **47,** 33, 38, 43 (1927).

[120] Fowler and Nordheim, Proc. Roy. Soc. **119,** 173 (1928); Oppenheimer, Phys. Rev. **31,** 66 (1928).

[121] Maxwell, Collected papers, Vol. 1, page 242.

field normal to the surface of the inner paraboloid at its vertex is given by $E = V/(p.\ln(c/p))$, where V is the difference of potential between the two paraboloids, p is the distance from focus to vertex for the inner paraboloid defined by $y^2 = 4px$, while c is the corresponding distance for the large one. If c is much larger than p so that the curvature of the outer paraboloid is negligible compared with that of the inner, the outer paraboloid may be

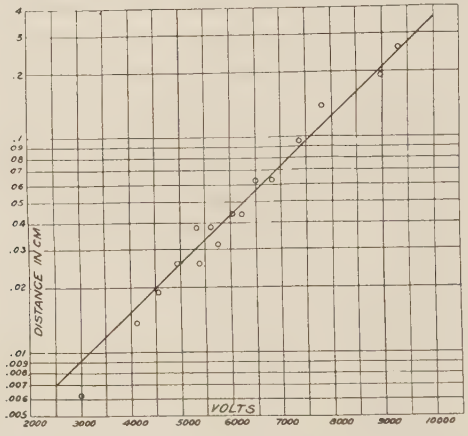


Fig. 7. Cold Cathode Emission. Relation of potential to distance between point and plane for constant current.

replaced without appreciable error by a plane surface. We thus have a means of calculating the potential gradient at the surface of a rounded point brought opposite a plane surface. This theory should, in fact, be better applicable to the case of a point and plane than the theory of confocal hyperboloids. If a is the radius of curvature at the point (at the vertex of the inner paraboloid) and c is the distance between the point and plane, then the equation becomes

$$E = \frac{2V}{a \cdot \ln(2c/a)}. \tag{38}$$

If the currents from the tungsten point which gave the data of Fig. 7 depended primarily upon the electric field at the surface of the point, then, since the current is kept constant, the field E must have remained constant as the distance was increased. According to Eq. (38) the logarithm of the distance c should then be a linear function of the voltage V. It is seen, in fact, that the experimental points agree within the experimental error with a straight line.

Because of the fact that the wire was broken off square instead of having the surface like that of a paraboloid, the actual field strength at the end of the wire must vary greatly at different parts of the surface and will not be given by Eq. (38). However, this equation must give a close approximation to the *average* field strength over the end of the wire. Thus since the *ratio* between

the fields at given points over the surface of the end of the wire will be approximately independent of the distance of the point from the plane, our conclusion that the voltage for a given field strength should vary linearly with the logarithm of the distance remains valid.

From the slope of the line in Fig. 7 we find that $dV/d(\ln c) = 1910$ volts and thus from Eq. (38), placing a equal to the radius of the wire (0.0108 cm) we find the *average* field strength E to be 350,000 volts cm^{-1}.

The foregoing results, together with those of Eyring, MacKeown and Millikan[116] prove that the field currents are actually a function of E only and not a function of the applied potential difference, V, and thus seem to disprove the conclusions of del Rosario[122] according to whom the current depended upon V and not E.

The following qualitative observations by Coolidge and Langmuir in connection with these experiments throw light on the nature of the field currents. When the currents exceed a few microamperes, spots of fluorescence appear on the glass and sometimes on the anode. These spots prove that the major part of the field currents emanate from minute spots on the cathode, for the fluorescence spots usually subtend angles of only a few degrees from their point of origin on the cathode, and sometimes less than 1°. In some experiments the cathode was a filament mounted on long leads. When the cathode was shaken the fluorescent spots always moved with it.

As the current and voltage are increased, the fluorescent spots become more brilliant and the walls of the tube become intensely heated at these points. This may lead even to the puncturing of the glass. These effects do not depend upon the presence of gas and the fluorescence spots are able to receive electrons continuously only because they become positively charged as a result of secondary electron emission.[123] As the field is increased, not only do the old spots become brighter but new ones continually appear. Thus experimental observations on the variation of field currents with field strengths should not agree accurately with the theory which led to Eq. (36) for this is deduced on the assumption that the current is uniformly distributed over the surface.

Field currents are obtained much more easily from surfaces covered with adsorbed films of electropositive materials. This effect was studied in some detail by Dushman, Langmuir and Hull in June and October, 1914.[124]

In the development of the high voltage kenetron, it was found necessary to degas the anodes by heating them to incandescence using voltages of 30,000 or 40,000 volts, the cathode temperature being reduced so that the current became limited by saturation emission, whereas in normal operation, the current is limited by space charge and the voltage drop from cathode to anode is relatively low. In 1915, kenetrons for 100,000 volts were manufactured. During exhaust the anode was made as much as 100,000 volts positive

[122] del Rosario, Jr. Frank. Inst. **203,** 243 (1927).

[123] Langmuir, Gen. Elec. Rev. **23,** 513 (1920).

[124] Unpublished work in Research Laboratory, Gen. Elec. Co., Schenectady; see also later work by Gossling, Phil. Mag. **1,** 609 (1926).

with respect to the cathode, while currents of 10 or 20 m.a. were drawn from the cathode. In this work the "cold cathode effect" became very noticeable at about 50,000 volts but by progressive degassing and heating of anode and cathode and baking the glass walls, it was finally possible to use 100,000 volts without appreciable "cold cathode effect". This effect when present, was evidenced by fluorescent spots on the glass, and by intense local heating at points on the anode which became white hot while the rest of the anode was below red heat. To avoid this effect entirely at 100,000 volts, it was found necessary to eliminate all sharp points from parts of the electrodes where the electric field was the strongest.

In 1914 some experiments were made to test the practicability of using thoriated tungsten filaments in kenetrons. At first it was not possible to obtain any higher electron emission than from a pure tungsten filament, but with continued degassing of the electrodes and the walls of the tube and by activating the filament by heating to 2100° it was possible to obtain the typical increased currents from the adsorbed film of thorium. Then, for the first time, it was found that at 30,000 volts, without any heating current through the filament, enough electron current was obtained from the thoriated filament to keep the anode red hot. At first the tube was not self-exciting, that is, if the anode was allowed to cool by removing the anode voltage for a couple of minutes, the tube would not start up again on applying the voltage. However ultimately, after further heat treatment, the tube was brought to the condition that it would start up as soon as 30,000 volts were applied to the anode, the currents being at first small but rising rapidly as the cathode became heated by the back radiation of the anode. These and many other similar experiences proved that the cold cathode effect occurred with an *activated* thoriated cathode at voltages of only one-third to one-half those needed to give a similar effect with pure tungsten. For this reason for many years much care was used to avoid all traces of thorium in both cathodes and anodes of kenetrons and Coolidge X-ray tubes.

In another experiment, a 5 mil V-shaped filament of tungsten containing 1.5 percent of ThO_2 was mounted in a tube with the tip of the V about 5 mm from a plate-shaped tungsten anode which had been degassed by heating white hot by electron bombardment. With the filament activated and giving currents of the order of 10^{-4} amp., the saturation current at 5000 volts was more than 10 times that obtained at 100 volts. A capsule containing metallic potassium was then broken in the tube in order to make it easier to maintain the thoriated filament in an active condition. With 7000 volts a magnificent blue fluorescence was then observed on the tungsten anode in spots and streaks and there was green fluorescence on the glass behind the anode. After the discharge had heated the anode red hot no further fluorescence on the anode was observed. With the filament heated to emit small currents at low voltages, the voltage was then raised to 10,000 and 20,000 volts. There were spots and streaks of green fluorescence on the glass which moved with the cathode when this was shaken. On a further rise of voltage sudden flashes or explosions seemed to occur within the bulb as though the charges on the glass

readjusted themselves. After each flash the positions of all the fluorescent spots were shifted and the intensity usually much increased. Soon, however, the spots returned to their original places and the intensity decreased. Raising the voltage to 35,000 a sudden flash occurred accompanied by a crackling sound and there then appeared a dazzling bright minute spot on the anode which rapidly heated the whole anode to incandescence. This spot moved over the surface of the anode when the cathode was moved proving that it was due to a fine pencil of cathode rays emanating from a spot on the cathode. Turning off the filament heating current, field currents of 5 ma at 30,000 volts were obtained with cold cathode. There could have been no appreciable positive ion bombardment of the filament for there was never any heating observed at the cathode. Momentary heating of the filament to 1800°K made the filament inactive so that no cold cathode emission was obtained at 30,000 volts, but at 35,000 volts again a similar flash occurred which rendered the cathode active.

Coolidge in 1914 and Coolidge and Langmuir in 1922 observed that when the field currents from a point in high vacuum were raised to several m.a., incandescent sparks were shot out from the cathode in the general direction of the anode. These particles described sharply curved paths, the curvature being convex on the side towards the anode, so that most of the particles did not strike the anode. Many of these particles when they struck the bulb rebounded; some were also observed to strike the disk-shaped anode and to rebound from its surface. This effect is probably caused by local heating resulting from the enormous current densities at the spot from which the electrons are emitted. Such particles are strongly negatively charged when they leave the cathode and are attracted toward the anode. Because of their high temperature, however, they emit electrons and become positively charged and are then repelled from the anode. It is difficult to understand, however, how positive charges of sufficient magnitude can be acquired to cause such marked curvature as that which was frequently observed.

The temperature rise, ΔT, caused by a current, i, (amperes) flowing out of a small spot in which the current density is I (amp. cm^{-2}) can be calculated to be approximately

$$\Delta T = \rho I i / 8\pi\lambda \tag{39}$$

where ρ is the specific resistance of the material (ohms cm) and λ is the heat conductivity (watts cm^{-1} deg^{-1}). Thus, for tungsten at room temperature $\rho = 5\ (10)^{-6}$ and $\lambda = 1.5$ so that the temperature rise is $\Delta T = 1.3\ (10)^{-7}\ I\ i$. Thus, even if the current density were $I = 10^9$ amp cm^{-2}, the rise in temperature at the emitting spot would only be 1.3° with an emission of 10 ma. The incandescent particles pulled out from the tungsten cathode therefore probably originate only at points on the surface where the surface metal is not in good thermal and electrical contact with the underlying material.. Possibly also the Bridgman effect caused an increase in the heating.[125]

[125] Bridgman, Proc. Nat. Acad. Sci. **10,** 297 (1921); Proc. Amer. Acad. of Arts and Sci. **57,** 131 (1922). Bridgman found that the resistance of gold and silver increased about one percent when the current density was 4 $(10)^6$ amp cm^{-2} and this increase in resistance was roughly proportional to the square of the current density.

Schottky has calculated[126] that the field at any surface irregularity is proportional to the average field over the surface, but about 10 times greater in magnitude provided these surface irregularities are small in comparison with the linear dimensions of the surface. This is in agreement with the fact that the observed values of E giving field currents are 10 or even 100 times less than those calculated either by Schottky's theory or by the newer theories involving wave mechanics.

The distribution of potential near the summit of a ridge-like elevation upon a surface may be roughly calculated as follows. Consider first the potential distribution in the neighborhood of an infinite conducting solid body which is bounded by two intersecting planes forming a dihedral angle β, this angle being less than π. Let α be the external angle, $\alpha = 2\pi - \beta$. Then it can be shown that the potential V at any point P in space outside the solid body is given by

$$V = Kr^n \sin n\theta \tag{40}$$

where

$$n = \pi/\alpha. \tag{41}$$

Here K is a constant, r is the distance of the point from the line of intersection O of the two planes and θ is the angle between the radius vector PO and its projection on one of the planes. The electric field E at the surface of either plane of the solid is given by

$$E = Knr^{n-1}. \tag{42}$$

We now consider a symmetrical ridge rising to a uniform height h above a horizontal plane and bounded by two plane surfaces forming an angle β, the bisecting plane being vertical. Let the line OC be a vertical line rising from any given point O at the summit of the ridge. At any given height c above O, the *field* due to the ridge (ignoring for the present the horizontal plane) is in a vertical direction and has the value

$$E = Knc^{n-1} \tag{43}$$

whereas the *field* due to the plane (ignoring that due to the ridge) is

$$E = E_0 \tag{44}$$

where E_0 is the field strength which would exist over the plane in the absence of the ridge. Similarly we may find the *potential* at a point at a distance c above O, first ignoring the plane and then ignoring the ridge, these potentials being respectively

$$V = Kc^n \tag{45}$$

and

$$V = E_0(h+c). \tag{46}$$

[126] Schottky, Zeits. f. Physik **14**, 63 (1928).

We may now eliminate c and K by assuming that the fields given by Eqs. (43, 44) are equal and the potentials given by Eqs. (45, 46) are equal.* We thus find for the field strength E at the surface of the ridge at a distance r from the vertex the approximate value

$$E = E_0 \left[\frac{nh}{(1-n)r}\right]^{1-n} \tag{47}$$

or

$$\log\left(\frac{r}{h}\right) = \log\left(\frac{n}{1-n}\right) - \frac{1}{1-n}\log\left(\frac{E}{E_0}\right) \tag{48}$$

Table VIII gives the values of r/h calculated by Eq. (48) for ridges having various summit angles β. The values of r/h are calculated for two values of E, viz., 10 E_0 and 100 E_0. Thus for the 90° ridge (each side at 45° slope) a field E varies inversely with the cube root of r, the distance from the vertex, and the field is more than 100 times E_0 for a distance of 2 $(10)^{-6} \times h$ from the vertex and is greater than 10 E_0 for a distance of 2 $(10)^{-3} \times h$.

TABLE VIII. *Field E at distance r from summit of ridge of height h on plane; β = angle at ridge, E_0 field at large distance from ridge.*

β	n	r/h	
		$E = 10E_0$	$E = 100E_0$
0	0.500	0.01	10^{-4}
30°	0.545	.0075	4.8×10^{-5}
60°	0.600	.0048	1.5×10^{-5}
90°	0.667	.0020	2×10^{-6}
120°	0.750	.0003	3×10^{-8}
150°	0.857	6×10^{-7}	6×10^{-14}
180°	1.000	0	0

These results may be used to calculate the field distribution over the square end of the wire of the experiment of Fig. 7. The actual wire may be approximately represented by a cylinder with rounded end on which is superimposed a square ridge with $\beta = 90°$ and $h = 1/10$ of the diameter of the wire, *i.e.*, $h = 0.0022$ cm. The average field E_0 over the surface was 350,000 volts cm^{-1}. The field was at least 10 times greater than this for a distance of 4.4 10^{-6} cm from the vertex of the ridge, giving a total area of 6 $(10)^{-7}$ cm^2. Since the current was 10^{-5} amp, the average current density over this area would have been 17 amp. cm^{-2}. A field of 100 E_0, *viz.*, 3.5 $(10)^7$ volt cm^{-1} extended for a distance $r = 4.4$ $(10)^{-9}$ cm according to the data of Table VIII, and to give the required current the current density would have been 1.4 $(10)^4$ amp. cm^{-2}.

Obviously we cannot expect to apply this theory when the value of r is less than the diameter of the atoms constituting the surface. Actually, the

* That this approximation gives reasonably accurate results follows from the fact that the field E at the summit of a hemispherical boss on a plane can be calculated in this way to be $4E_0$, whereas a rigorous calculation, which may be made for this case, gives $E = 3E_0$.

emission of fine pencils of cathode rays proves that the emission cannot occur from a ridge but from the serrated points along such a ridge. The theory, however, aids in forming a conception of the magnitude of the effects to be expected from surface irregularities.

Some experiments under unusually good vacuum conditions were carried out by W. D. Coolidge in 1923 (unpublished work). Two short V-shaped filaments were mounted in a small bulb so that the tips of the V's came within about 0.1 mm. One of these filaments was one mil diameter (0.0025 cm) and the other one was 8 mil diameter (0.020 cm). The tube was well exhausted and the filaments were degassed at very high temperature. During much of the work with the tube, it was immersed completely in liquid air to improve the vacuum still further.

The following are some of Coolidge's observations. With the bulb in liquid air after the filament has been heated to high temperature and allowed to cool, the filament is very inactive. If the voltage is gradually raised to the neighborhood of 10,000 volts until an appreciable field current can be measured on the galvanometer and the voltage is then kept constant, the current gradually increases at a rapidly accelerating rate but suddenly becomes constant and remains so. The filament is then in an "active" condition. The filament remains active if the voltages are taken off for a long time and reapplied. If the filament is heated to 1600° for 10 or 15 seconds, or to 2400° for one second, the filament becomes inactive but can be made active again by applying a sufficiently high voltage until the current again rises to its limiting value. With continued experimentation, the voltage required to activate the filament gradually increases. Heating the anode filament (the large filament) even to 2500° while the bulb is in liquid air has no effect on the field currents obtained from the small filament.

After the liquid air is removed, as the bulb temperature rises, the field currents at a given voltage first decrease and then, as the temperature approaches room temperature, the currents increase greatly. Heating the filament above 1600°K still caused a great decrease in activity.

In some experiments oxygen at a pressure of a few microns was introduced. Usually this caused no marked increase in activity if the bulb was in liquid air or at room temperature, but if the bulb was then heated moderately the filament became very active. This suggests that traces of alkali metal escaped from the glass and are held by a monatomic oxygen film on the tungsten. Heating the filament above about 1600° drives off some of the oxygen from the tungsten surface and renders the film inactive.

In general, the filament could be rendered active by raising the voltage sufficiently to draw fairly large currents. This activation was often accompanied by a crackling sound suggesting that disruptive discharges over the surface of the glass caused gas evolution which altered the cathode in such a way as to make it active. Activation also occurred as mentioned above by heating the bulb after oxygen treatment. Introduction of hydrogen seemed to have no effect on the activity. The cathode was rendered insensitive by heating to 1600 or 1800°K and sometimes by drawing excessively large currents (a few milliamperes).

The difference in the voltages necessary to draw currents from the inactive and from the active filaments frequently differed in a ratio as great as 10 to 1 or even 20 to 1.

These results emphasize the importance in future work on field currents, of exceptional care in controlling surface conditions of the cathode. After each measuring of field current it would seem desirable to test the condition of the filament by measuring its electron emission at a suitable testing temperature.

Finally, attention may be called to two very recent papers[127] dealing on the basis of wave mechanics with field currents, and especially with the modification introduced into Eq. (36) by the presence of a thin adsorbed layer of electropositive material. Comparison with experiments indicates a possibility of satisfactory interpretation of complications due to surface layers and indicates stable films of thicknesses of the order of 2.5 $(10)^{-8}$ cm.

(7) **Electron emission due to electron bombardment** is one of several phenomena which have been described as *secondary electron emission*, although it was originally called *delta ray* emission.[128] It is impossible to tell whether a secondary electron is simply a reflected primary, or whether it is an electron, originally of the metal, which has been ejected by the primary. For all practical purposes these two possibilities need not be differentiated. The fact however, that the number of secondaries sometimes exceeds the number of primaries proves that real emission does occur.

The main facts of this secondary emission may be summed up as follows: (a) The secondary emission depends considerably on the surface treatment of the metal, being generally reduced by heat treatment and degassing and increased by contamination, especially with electropositive impurities. (b) The emission increases from a low value (perhaps zero) for zero velocity of the primary electrons, rises to a maximum for primary electrons of a few hundred volts energy, and then slowly diminishes with further increase of energy of the primaries. (c) The maximum ratio of the *number of secondary electrons per primary electron* is between about 1.0 and 1.5 for well degassed ordinary metals; it may reach values of 3 or 4 for metals not specially treated. For electropositive metals this ratio is greater. For films of alkali metal on an oxidized metal surface, the number of secondaries per primary may be as large as 8 or 10. (d) Although the general trend of secondary emission with voltage is as described, there are superimposed small variations of the nature of subsidiary maxima which suggest the existence of *critical potentials* at which secondary emission by new processes set in. (e) The *velocities of the secondary* electrons are, for the most part, quite small, amounting to only a few volts even for primary electrons of a thousand volts. (f) The *directions of emission* of the secondary electrons appear to be distributed more or less at random. (g) Except for its effect in conditioning the surface, moderate *temperature variations* of the surface appear to have little effect. (There is

[127] Stern, Gossling and Fowler, Proc. Roy. Soc. **A124,** 699 (1929); Stern, Proc. Camb. Phil. Soc. **25,** 454 (1929).

[128] Campbell, Phil. Mag. **22,** 276 (1911); **25,** 803 (1913); **28,** 286 (1914); **29,** 369 (1915).

indication, however,[129] that the critical potentials mentioned under (d) are strongly emphasized if the surface is extraordinarily well freed from adsorbed gases, as by keeping it continually hot in a good vacuum.) The foregoing characteristics have been established by many observers.[130] A number of these agree in finding indication of about 0.50 of incident electrons of very low velocity reflected from ordinary metal surfaces, though the interpretation of results is particularly uncertain for very low velocities.

In view of the importance of reflecting power in the theory of thermionic emission (see section B 1), it is of interest to note the following strong evidence *against* a reflecting power of the order of 50% at low velocities. Langmuir and Jones[131] in their studies of ionization phenomena between an axial tungsten cathode in a cylindrical molybdenum collector measured the fraction of the primary electrons which left the cathode that were collected by the cylinder without having lost energy by inelastic collisions (see also section D 2b). If the logarithm of this fraction is plotted as ordinate against the pressure, straight lines are obtained whose slope is R/λ_1 where R is the radius of the cylinder and λ_1 is the free path of the electrons for inelastic collisions at unit pressure.

The intercept of this line gives the fraction of the electrons that would be collected if there were no collisions. This should be unity if it were not for the reflection of electrons incident on the cylinder. Actually, the intercepts that are found experimentally range from 0.77 to 0.83 and are independent of the nature of the six different gases used. This result proves that with molybdenum surfaces which have been cleaned by bombardment by positive ions the reflection coefficient, for electrons striking the surface normally with energies corresponding to 10 to 20 volts, lies between 0.17 and 0.23.

The shapes of the curves obtained by Langmuir and Jones were in good agreement with the empirical equation

$$1/(1-r) = 1+0.0645V^{1/2}$$

where r is the reflection coefficient of electrons having an energy corresponding to V volts. Their data are wholly inconsistent with any increase in r as V decreases. The results thus indicate that the reflection coefficient becomes very small for low velocity electrons.

The electric image theory of the forces acting on an electron close to the surface of the metal also gives strong reasons for the view that the reflection coefficient should approach zero as the velocity of the incident electrons becomes very small. Consider, for example, electrons in equilibrium with a surface at 2000° which impinge on the tungsten surface. These electrons will

[129] Krefft, Phys. Rev. **31,** 199 (1928).

[130] Gehrts, Ann. d. Physik **36,** 995 (1911); Hull, Phys. Rev. **7,** 1, 141 (1926); McAllister, *ibid.* **21,** 122 (1923); Farnsworth, *ibid.* **20,** 358 (1922); **25,** 41 (1925); Petry, *ibid.* **26,** 346 (1925); **28,** 362 (1926); Gill, Phil. Mag. **45,** 864 (1923); Horton and Davies, Roy. Soc. Proc. **97,** 23 (1920); Phil. Mag. **46,** 129 (1923); Stuhlman, Phys. Rev. **25,** 234 (1925). Tate, *ibid.* **17,** 394 (1921).

[131] Langmuir and Jones, Phys. Rev. **31,** 401 (1928). See particularly Fig. 15.

have, on an average, energies of about 0.2 volt, but as they approach the metal they are acted on by the attractive force corresponding to the electric image. If the work function, 4.5 volts, is entirely due to the electric image force, the electrons will have an energy of about 4.7 volts by the time they reach the actual metal surface. If they should lose only 0.2 of a volt or about 4 percent of their energy, they will be unable to escape from the surface. According to some modern theories involving wave mechanics, the electric image force may be equivalent to energies of 10 or even 15 volts which is partly balanced by the electron pressure within the metal. On this basis, energy losses of much less than 4 percent would prevent the reflection of the electrons.

It was pointed out by Langmuir[132] that from the heat conductivity of a metal the time of relaxation of the electrons can be calculated. For tungsten this time is of the order of 10^{-15} seconds. This is about the time that it takes for an electron, having unusually high energy, to lose all but $1/\epsilon$th of its excess energy. A 0.2 volt electron which has an energy of 4.7 volts when it reaches the metal, is traveling with a velocity of 1.2×10^{8} sec. and therefore when it penetrates within the metal to a distance of only a single atomic diameter, remains in the metal about 0.4×10^{-15} sec. and thus should lose about half of its energy because of the normal heat conductivity of the metal. Thus only an extremely small part of the electrons should lose less than 4 percent and only these can be reflected from the surface.

There is another effect, however, which must also greatly decrease the probability that low velocity electrons will be reflected. When an impinging electron has penetrated within the surface layer of atoms, all other free electrons are repelled from it, leaving a slightly greater positive charge than before the incident electron arrives. This is a kind of internal image force analogous to that postulated by Debye and Hückel in their theory of electrolytes. The length of time it takes to bring about this redistribution of charge can be estimated roughly as follows. In a metal of specific resistance ρ consider a charge e brought to the center of a sphere. Let t be the time that it takes for the charge within the sphere to fall to $1/\epsilon$th of its original value. We thus find $t=10^{9}\rho/4\pi c^{2}$ where c is the velocity of light. For tungsten $\rho=5\times10^{-6}$ and thus we find $t=4\times10^{-19}$ sec. The redistribution, therefore, within the metal occurs in 1/1000 of the time of transit of the electron. At some slight distance from the surface where the electron concentration is lower, the time of relaxation will be comparable with the rate of passage of the electron and there will be relatively large hysteresis effects which will lead to considerable energy dissipation and thus tend to prevent reflection.

Besides being of importance in certain very special types of discharge (Lilienfeld X-ray tube,[133] Hull dynatron[134]) and in equilibrium theories of thermionic and photo-electric emission, electron reflection has recently become of great interest in the interpretation of the electron itself and the foundations of the quantum theory. This latter aspect developed from

[132] I. Langmuir, Phys. Rev. **8**, 171 (1916).
[133] Lilienfeld, Ber. d. König. Saechs. Ges. d. Wiss. **66**, 76 (1914).
[134] Hull, Phys. Rev. **7**, 141 (1916); Proc. Inst. Rad. Eng. **6**, 5 (1928).

studies six or eight years ago by Davisson and Kunsman[135] of the angular distribution of that very small fraction of rather fast primary electrons incident on a metal surface which are reflected with practically the entire original energy. Since then the phenomenon has been carefully investigated by Davisson and Germer[136] using reflection from the faces of nickel crystals. These experiments show strong reflection at particular angles of incidence and reflection which are related by the ordinary laws of diffraction of waves by a crystal lattice, and prove that the electrons behave like trains of waves of wave-length $\lambda = h/p$, where h is Planck's constant and p is the momentum of the electron. Furthermore it is found that the refractive index of the crystal for these waves is not unity, but that the crystal possesses a sort of "dispersion" curve for electrons of various equivalent wave-lengths. This "wave nature" of electrons is strikingly confirmed by the experiments of G. P. Thomson, who has shown that electrons in passing through thin films of metal, or when incident at grazing angle on a sputtered film on quartz, are scattered as if diffracted characteristically for waves of length $\lambda = h/p$ by the atomic lattice grating.[137]

Under the influence of electron bombardment secondary electrons may be emitted not only from metallic surfaces but from the *surface of insulators*. High velocity electrons striking glass surfaces may cause more than one secondary electron to be emitted per incident electron. Thus the glass surface becomes positively instead of negatively charged. In most high vacuum, high voltage discharge tubes, especially when operated at higher voltages than normal, spots of blue or green fluorescence are seen on the glass which owe their origin to pencils of cathode rays pulled out from the cathode by the intense fields. There is marked heat evolution at such fluorescence spots proving that these spots are at relatively high positive voltages with respect to the cathode. Since these phenomena occur in tubes in which no appreciable current can be carried by positive ions, it follows that the positive charge is maintained by secondary electron emission.

Langmuir[137a] has described experimental studies of these effects of secondary electron emission from glass surfaces. In general, two types of discharge are possible. First, those in which the walls become negatively charged and thus repel all further electrons except those few necessary to balance the small positive ion current. Fluorescence under these conditions, in presence of traces of gas, may be produced by positive ion bombardment. There are some indications that the red fluorescence sometimes observed with low pressures of oxygen is due to such positive ion bombardment. In the second type of discharge, secondary electron emission keeps the walls positive so that they continue to receive electrons, and a large part of the energy of the discharge may be then liberated in the form of heat at the walls.

[135] Davisson and Kunsman, Science, **64,** 522 (1921); Phys. Rev. **22,** 242 (1923).

[136] Davisson and Germer, Phys. Rev. **30,** 705 (1927); Proc. Nat. Acad. Sc. **14,** 317 (1928); Davisson, Jour. Frank. Inst. **205,** 597 (1928).

[137] G. P. Thomson, Nature **122,** 279 (1928); Proc. Roy. Soc. **A117,** 600 (1928); **119,** 651 (1928); **125,** 352 (1929).

[137a] Langmuir, Gen. Elec. Rev. **23,** 513 (1920).

A repetition[137a] of Lilienfeld's experiments[138] proves that his high vacuum discharges through long glass tubes are dependent entirely upon secondary electron emission from the walls of the tube.

(8) **Electron emission due to impact of metastable atoms** was for some years suspected of producing a part of the electron emission from a cathode, in the presence of an excited gas, which had at first been ascribed entirely to photo-electric action of the radiation. It was first definitely proved by Webb[139] and Miss Messenger,[140] who interposed screens of quartz and fluorite of measured optical transparency between the excited gas and electron emitting electrode, and from the amount of diminution of emission they were able to evaluate the relative importance of emission due to impact of metastable atoms and due to light in various parts of the spectrum. They showed that the effect of metastable atoms is comparable with, and in some conditions considerably more effective, than photo-electric action in liberating electrons.

Recently Oliphant[141] has shown that the beam of positive ions of helium, which are shot through an orifice in a negatively charged Langmuir plane collector, may be largely transformed into a beam of metastable helium atoms of similar speed by allowing the ions to impinge at a glancing angle upon a metal surface. These metastable atoms, in turn, were found to liberate electrons upon impact with a second metal surface. The probability of liberation was at least several percent, and the simplest interpretation of the results is given by supposing that every metastable atom which reverted to the normal state at a metal surface set free one electron. Also a certain percentage of these metastable atoms was reflected at the surface, this percentage decreasing with depth of penetration, *i.e.*, with speed and with glancing angle, and increasing as the surface was heated. The amount of reflection was generally beween 10 and 50 percent. The electrons liberated by these metastable atoms possessed kinetic energies lying between a maximum value of $V_0-\phi$ (where V_0 is the energy of the metastable state and ϕ is the work function of the metal) and a minimum value of about 2 volts, which may have significance in connection with Sommerfeld's theory of metals. Apparently the energy of emission of the electrons had no relation to the kinetic energy of the impinging metastable atoms.

Very significant evidence regarding the secondary emission of electrons from a negative cold electrode in a discharge has recently been obtained by Uyterhoeven and Harrington.[142] They introduced into a region of uniformly ionized neon gas, whose ionization was maintained by supplementary electrodes, the system of electrodes shown in Fig. 8. A circular disk electrode D was surrounded by an ample guard ring G and was pierced at its center by a pinhole. Close behind this pinhole was a Faraday box F. Between D and G

[138] Ann. d. Physik **32,** 673 (1910); *ibid.* **53,** 24 (1914); Leipziger Ber. **63,** 34 (1911).

[139] Webb, Phys. Rev. **24,** 113 (1924).

[140] Messenger, Phys. Rec. **28,** 962 (1926).

[141] Oliphant, Proc. Roy. Soc. **A124,** 228 (1929).

[142] Uyterhoeven and Harrington, Phys. Rev. **35,** 124 (1930); also more recent unpublished work.

was a narrow open gap as shown. There was no communication between the main discharge and the region below these electrodes except through this gap and the pinhole. Facing D was another Faraday box B, exclosed by glass except on the open end facing D. This box B could be moved vertically to various distances from D. The experiments which are of immediate significance are the following:

(1) Whatever the potential of the electrodes DG with respect to the ionized gas above them, there was always a considerable current between F and G, *which would pass in either direction* according to the sign of the field between F and G, and which was quite independent of the potential of DG with respect to the gas. Furthermore, applying a field between D and G had very little effect on this current. These observations proved that the conductivity be-

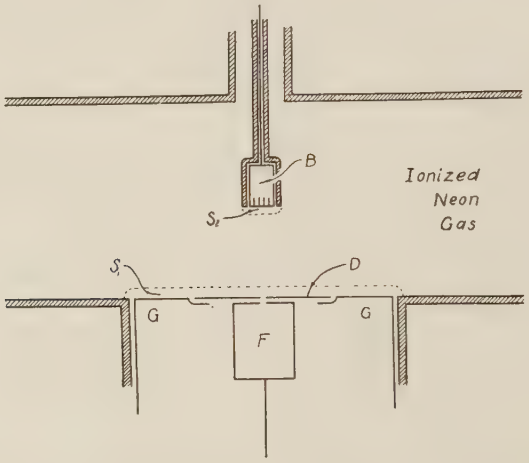


Fig. 8. Apparatus of Uyterhoeven and Harrington for study of electron emission due to positive ions and metastable atoms. Nickel electrodes.

tween F and G was not due to charged particles passing from the main discharge into the region below DG. The currents, moreover, were of such a magnitude as probably to exclude their explanation as of photo-electric origin, especially in view of the complicated reflection which would have been involved in producing photo-electric emission from the under side of G. The only reasonable explanation of the currents between F and G thus seems to be action of long-lived metastable atoms which had diffused through the gap.

(2) The collector B was used to prove the existence of a very considerable electron emission from the upper surface of the disk D which was exposed to the main discharge. To do this, the collector B was made negative with respect to the surrounding space by some definite amount, say 100 volts. It was then surrounded by a positive space charge sheath S_2 and it collected ions but not electrons from the surrounding ionized gas. Keeping this condition constant, the potential V of the electrodes DG was now varied from a small to a large negative value. So long as V was less negative than 100 volts, the elec-

trodes DG were practically without influence upon the current to B. Just as V was raised to 100 volts, a considerable negative current was received by B. Making V still more negative had relatively little further effect. This negative current to B must have been an electron current proceeding from D, and the fact that it came in definitely just when the potential of D became more negative than B proves that these electrons must have originated at the surface of D itself. If they had originated within the space charge sheath S_1, rather than at the electrode D, their velocity distribution would not have been so homogeneous.

Now these electrons may have been emitted by the action of light, of metastable atoms or of positive ions. Without doubt all three processes contributed. But the following reasons indicate that the metastable atoms exerted by far the preponderating effect. The amount of this electron current density was about half the total current density to D, indicating that approximately half the current to D was due to these secondary electrons and the other half to the incoming positive ions. The evidence in the following section (B 9) is that electron emission due to positive ion bombardment is far less effective than this. The same conclusion may be drawn from the fact that the results were essentially the same whether the potential of B was 50, 100 or 200 volts. Similarly no photo-electric effects of this order of magnitude have ever been recorded. Consequently we seem left with action of metastable atoms as the only adequate explanation.

This work is still in progress, obtaining additional data for a variety of gases, and also measuring electron free paths and the fraction of the current to the cathode which is carried by positive ions.

Some earlier work by Uyterhoeven[143] with apparatus similar to parts D and G of Fig. 8 led to results in He, Ne and A which, in the main, can be similarly explained, and whose significance in regard to electron emission due to positive ion impact is discussed in the following section (B 9). These conclusions are substantiated by Found,[143] who has used the characteristics of a cylindrical collecting electrode to evaluate the importance of emission of electrons by action of metastable neon atoms in the positive column of a neon discharge. The positive ion current to the negative collector is proportional to the area of the surrounding sheath, whereas the electron emission from it due to metastable atoms depends only on the area of the collector itself and hence is constant. In a positive column of 5 cm diameter in neon at pressure 0.225 mm, operating at 0.4 amp., the secondary electron emission from a negative tungsten collector is 7.5 milliamps/cm^2 compared to a positive ion random current density of 0.82 milliamps/cm^2. The current to a 2 mil collector at 100 volts negative to the discharge was thus approximately two-thirds due to secondary emission and one-third due to positive ions.

(9) **Electron emission due to positive ion bombardment** is another phenomenon which has long been suspected, but only recently proved, and really very little is known about it. Practically all of the older work is vitiated by unrecognized complicating effects.

[143] Uyterhoeven, Proc. Nat. Acad. Sc. **15,** 32 (1929); also thesis for Cal. Inst. Tech. 1929; Found, Phys. Rev. **34,** 1625 (1929).

The simplest conditions obtain when a metal surface is bombarded by positive ions in a high vacuum, and the resulting electron emission is measured after suitably allowing for positive ion reflection. Klein[144] and Jackson[145] have done this, using heated strips emitting positive ions of alkali metals as ion sources. The results of the two authors are in serious disagreement, due, we believe, to an unrecognized complication arising from the experimental arrangement in the former case. Jackson found, for K ions impinging perpendicularly on Al, Ni and Mo, no detectable electron emission (*i.e.*, less than 0.5 percent) for ion velocities below 200 volts for Al, 300 volts for Ni, 600 volts for Mo after degassing of these metals by heat treatment. Above these velocities the electron emission increased regularly to 7.0% for Al, 4.2% for Ni and 3.8% for Mo at 1000 volt velocities. Without the heat treatment the emission was detectable at about half the above minimum velocities and was about double the above values at 1000 volts. Positive ions of Cs were slightly more, and those of Na and Rb somewhat less effective than those of K. The electrons thus emitted were very slow, practically all of them being stopped by a retarding field of one volt.[146]

In the above experiments not more than 4% of the incident *ions were reflected*. At more glancing angles, however, and particularly with degassed surfaces, the percentage of reflection may be much greater, and the reflected ions possess a complicated distribution of directions and of retained energies.[147]

A very decisive test of the effectiveness of alkali ions in producing secondary electron emission from bombarded metal surfaces can be made by utilizing the remarkable constancy of the ion currents from a sufficiently hot filament in contact with the alkali metal vapors, since such currents depend only on the rate at which vapor atoms strike the filament and are entirely independent of applied voltage, provided this is sufficient to overcome space charge.[148] It is observed that if such a current of caesium ions from a filament to a surrounding axial cylinder is saturated at 50 volts, a sudden increase of voltage to 250 volts causes no perceptible change in the current. This proves that there is no change as large as a fraction of one per cent in the secondary electron emission in the range 50–250 volts, which makes it extremely probable that the total secondary electron emission itself is negligible.

It is particularly important to know the magnitude of this phenomenon in mercury vapor. Langmuir and Mott-Smith[149] found that the current to a negative nickel collector bombarded by Hg ions showed no evidence of secondary electron emission due to bombardment up to 1160 volts, though their conditions were such that an emission of not over 10% would not have

[144] Klein, Phys. Rev. **26,** 800 (1925).

[145] Jackson, Phys. Rev. **28,** 524 (1926); **30,** 473 (1927).

[146] See also Baerwald, Ann. d. Physik **60,** 26 (1919).

[147] Read, Phys. Rev. **31,** 155 (1928); Gurney, *ibid.* **32,** 467 (1928).

[148] Langmuir and Kingdon, Science **57,** 58 (1923); Killian, Phys. Rev. **27,** 578 (1926). Note particularly the flat saturation curves in Figs. 5, 6, p. 585 of Killian's paper.

[149] Langmuir and Mott-Smith, G. E. Rev. **27,** 544 (1924).

been detected on account of uncertainties in the "edge correction" to the currents to the collector. Then Dällenbach, Gerecke and Stoll[150] showed that Hg^+ ions incident on an iron electrode with energies up to 3000 volts are not more than perhaps 1 percent effective in liberating electrons from the electrode.

In some recent unpublished work, Langmuir and Sweetser have measured the variation of current to a negative collector with voltage far more accurately than in the work of Langmuir and Mott-Smith. The collector was of molybdenum of the guard-ring type. The outer diameter of the guard-ring was 1.92 cm while the diameter of the inner part of the collector was 0.64 cm, the separation between the two collectors being less than 0.005 cm. Both parts of the collector were always kept at the same voltage, but the current was measured only for the central part. The fractional variation of current per volt change in collector potential, which we may represent by β, was measured under a wide range of conditions. The largest observed variation ($\beta = 0.001$) was found only under conditions where the sheath thickness (x) was very large (0.20 cm). By varying x by altering the intensity of ionization, for example by changing the cathode current, it was found that β varied with x according to the equation

$$\beta = \beta_0 + 1.65x^2/(Vr_0^2) \tag{49}$$

where V is the negative voltage on the collector and r_0 is the radius of the outer perimeter of the guard-ring. In this equation the term involving x^2 is of the form to be expected for a residual edge correction due to an insufficiently large guard-ring. This equation was found to express the results for discharges for both argon and mercury vapor.

With a greater intensity of ionization the term involving x^2 becomes quite small as compared with β_0 and thus the value of β_0 could be found with reasonable accuracy. For the discharges in mercury vapor $\beta_0 = 0.0002$ volt^{-1}. For argon $\beta_0 = 0.0006$ volt^{-1}. In these experiments the voltage of the collector was varied from 50 to 150 volts. In this whole range of 100 volts, therefore, it may be concluded that if there is any electron emission produced by positive ion bombardment it varied by about 2 percent for mercury and 6 percent for argon, these being expressed as percentages of the total positive ion current flowing. (Experiments with collectors designed to catch the electrons emitted from the surface of a negatively charged collector are now in progress by Tonks and Sweetser).

Similar experiments by Uyterhoeven,[143] at rather lower pressures, do not agree very accurately with Eq. (49), but if this equation is used as a rough approximation then values of β_0 may be computed. Typical values are: He–Ni, 0.0020; A–Ni, 0.00095; A–C, 0.0012.

It should be noted that these are "apparent" values, based on the assumption that the total collector current is a positive ion current. They would be raised if part of this current is due to emitted electrons,—doubled for example

[150] Dällenbach, Gerecke and Stoll, Phys. Zeits. **26,** 17 (1925).

if half the current is due to electrons as in the case of Ne–Ni investigated by Uyterhoeven and Harrington.[142]

The most direct investigation of electron emission due to positive ions of a gas discharge is by Penning,[151] who projected neon ions, from a discharge, through a tube and allowed them to strike a copper, silver or iron electrode after the desired speed regulation in an auxiliary field. The resulting probability of electron emission as a function of the ion energy is shown by Fig. 9. It is interesting to note the emission even at zero velocity of impact. Extrapolation of the experimental curve gives the probability at zero velocity somewhere between 2 and 2.5%. Penning suggests that this will be found

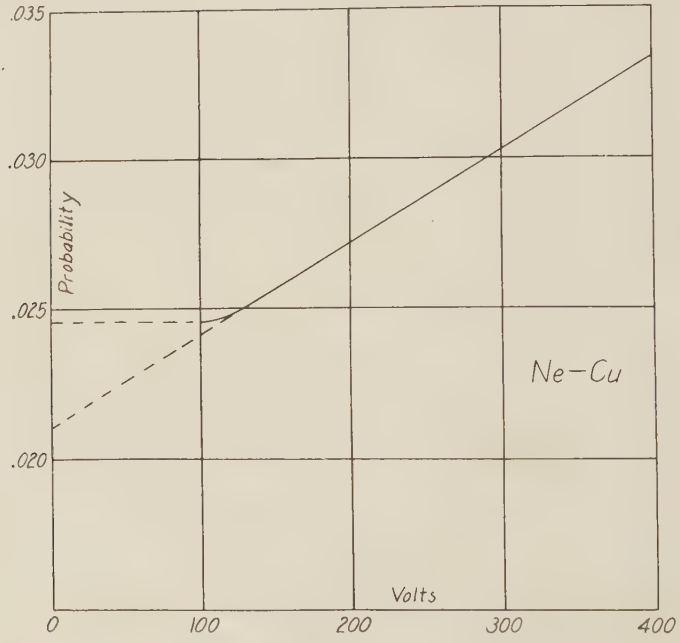


Fig. 9. Probability of electron emission by positive ion bombardment at V volts. (Neon ions, copper electrode.)

whenever the ionizing potential of the gas atom exceeds twice the work function of the electrode, $V_i > 2\phi$, because the ion must possess enough energy to extract an electron to neutralize itself, and also another electron. Oliphant[141] points out that Penning's results probably include a contribution from metastable atoms produced as in his own work. Nevertheless the increase of emission with ion velocity indicates that Penning has observed a true electron emission due to slow speed ions, even though his values may be too high on account of the complication mentioned by Oliphant.

(It is perhaps significant, in support of Penning's views, that cold cathode discharges in Hg vapor require a surprisingly high voltage in comparison with

[151] Penning, Vers. K. Ak. Amsterdam **36** (1927); Physica, **8,** 13 (1928); Proc. Amst. Acad. **31,** 14 (1928).

other gases whose ionization is known to be more difficult. This may be because the inequality $V_i > 2\phi$ is considerably less pronounced in the case of Hg than in the case of the gases in question, so that the positive ions are less effective in ejecting electrons from the cathode. If this be true, discharges in K vapor should be still more difficult, provided the cathode is of a material like oxygen-free zinc which will not adsorb K atoms or ions and thereby lower its work function.)

Finally, the yield of electrons liberated from cathodes by positive ion impact has been estimated indirectly and on questionable grounds by Klemperer,[152] from data on the minimum sparking voltage between parallel metal planes, combined with values of the "*effective*" (not minimum) ionizing potentials of the gases.[153] These estimates involve the assumption that every electron ionizes *every* time it falls through this "effective" ionizing potential, this being the characteristic of and the explanation of the minimum sparking voltage (an assumption and a statement which are not obvious). This estimate agrees in order of magnitude with calculations based on the Holst and Oosterhuis theory of sparking (see Part II). Klemperer's values of percentage yield of electrons per positive ion impact are given in Table IX.

TABLE IX. *Yield of secondary electrons per positive ion.* (*Estimated indirectly.*)

gas	air	CO_2	H_2	A
Yield	0.66%	0.09%	0.74%	0.7%
"*u*"	5.8	8.4	2.4	3

"*u*" = kilovolts/cm. ÷ pressure in cm Hg and is a measure of the mean energy acquired by the ions in passing through the gas between the plane parallel electrodes.

(10) **Emission due to chemical action** probably plays a very minor role in gas discharges. The phenomena are varied in nature. For example, air, especially when moist, when in contact with phosphorous is charged with both positive and negative ions of such low mobility that they obviously contain very complex products of the chemical reaction. Hot platinum in contact with phosphorous vapor emits positive ions but no appreciable negative ions or electrons. Gaseous products of electrolysis are ionized, though this is probably due to bubbling through the electrolyte rather than to chemical action.

Electrons are emitted by chemical action of a number of gases on such electropositive metals as sodium, potassium, amalgamated aluminum. Among the gases investigated are H_2S, HCl, CO_2, H_2O, $COCl_2$, $CSCl_2$, O_2, Cl_2, Br_2. Cases of emission of as many as one electron for every 1600 reacting molecules have been observed.[154] The emitted electrons have velocities of the same ordre as those ordinarily observed in photo-electric emission, but with that absence of saturation which we have shown to be characteristic of a composite surface.

[152] Klemperer, Zeits. f. Physik **52**, 650 (1928).

[153] Lehmann and Osgood, Proc. Roy. Soc. **A115**, 609 (1927); Lehmann, *ibid.* **115**, 624 (1927).

[154] Haber and Just, Ann. d. Physik **30**, 411 (1909); **36**, 308 (1911); Zeits. f. Elektrochem. **16**, 275 (1910).

Doubtless some cases of emission ascribed to chemical action are really thermionic emission arising from high temperatures produced by the reaction, while others may be of photo-electric origin associated with chemiluminescence. There appears to be no reason, however, for excluding electron emission by chemical action as a real phenomenon, since activated molecules are known to exist as intermediate products in chemical reactions,[155] and these activated molecules should be able to cause electron emission if their energy is sufficient, just as such emission occurs at contact of metastable atoms with electrodes.

An excellent discussion of ionization and chemical action is given by Richardson in his "Emission of Electricity from Hot Bodies."[156]

C. Disappearance of Electrons and Ions

Electrons and ions may be lost by any given element of volume of ionized gas by recombination, by absorption or by neutralization at the surfaces of electrodes, or by motion away from this element by volume. This latter process will be discussed in the following section entitled, "The Motions of Electrons and Ions."

(1) **Disappearance of electrons and ions at the electrodes and walls** involves, of course, their motion to the electrodes, which will be treated later. In addition to this, there are certain phenomena characteristic of the actual absorption or neutralization process. Of these phenomena, reflection and secondary emission of new charged particles have already been discussed (B 7, 8, 9).

Absorption of electrons by electrodes involves heating of the electrodes by the amount $e(\phi+\bar{V})-2kT$ per electron,[157] where $\bar{V}$ is the average energy of the energy of the incident electrons (in equivalent volts) and T is the excess in temperature of the electrode above its external connection. $\bar{V}$ does not include the energy gained close to the electrode surface as the result of surface forces, since this is included in ϕ. The term $2kT$ (which amounts only to e times 1 volt for $T=5886°$) is generally negligible. If the electrons reach the electrode in a retarding field, so that secondary emission and reflection of electrons occur, these phenomena alter the heating in an obvious manner, the net heating per incident primary electron being

$$H=e(\phi+\bar{V})-n_s e(\phi+\bar{V}_s)-(1-n_s)2kT \tag{50}$$

where n_s is the number of secondary electrons per primary, and $\bar{V}_s$ is the average initial energy of the emitted secondary electrons. If the incident energy $\bar{V}$ is of the order of 10 volts, the term $n_s e(\phi+\bar{V}_s)$ is probably less than 10% of the term $e(\phi+)\bar{V}$. At higher incident energies n_s increases (up to a maximum when $\bar{V}$ is of the order of 300 volts), but $\bar{V}_s$ remains so small relative to $\bar{V}$ that it is very doubtful if the cooling effect of the second term can

[155] Smyth, Proc. Nat. Acad. Sc. **11**, 679 (1925).

[156] Richardson, "Emission of Electricity from Hot Bodies," Chap. IX.

[157] Richardson and Cooke, Phil. Mag. **20**, 73 (1910); **21**, 404 (1911), with the correction of $2kT$ in place of $(3/2)kT$.

ever equal the heating effect of the first term. The more important terms of Eq. (50) have been tested both for electrodes in *vacua*[157] and in ionized gases.[158]

Absorption or neutralization of positive ions at electrodes involves more complicated energy relations. Consider the cycle of Fig. 10. A positive ion of

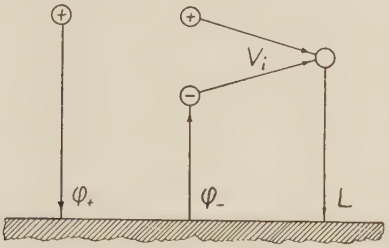


Fig. 10. Heat of neutralization ϕ_+ of a positive ion.

negligible kinetic energy reaches the surface, is neutralized, and liberates energy $e\phi_+$. Or, alternatively, an electron escapes from the surface absorbing energy $e\phi_-$, unites with the ion liberating energy eV_i, and the neutral atom *may* remain on the surface liberating the heat of absorption eL of the neutral atom. Comparing these equivalent processes, we get

$$\phi_+ = V_i - \phi_- + L, \tag{51}$$

which is the relation originally proposed by Schottky, except for the inclusion of the term L (which vanishes if the ion-atom is not permanently deposited on the electrode).

While this equation is correct as regards total energy, it is not necessarily correct as regards *heat gained by the electrode*, because a part of the energy V_i may be radiated away in the process of combination. The process of neutralization may *actually* be somewhat similar to that shown on the right in Fig. 10, since, when the ion has approached within a distance of about ten atomic diameters of the surface it attracts electrons with a force sufficient to extract them from the surface. The subsequent time before the ion strikes the surface is of the order of 10^{-12} sec., and therefore much smaller than the average duration τ of a free atom in an excited state (A 3). Nevertheless, at or near the surface the combining ion must part with energy V_i, and it is quite possible that a fraction τ of this is radiated. Perhaps this radiated energy accounts for the faint continuous glow often observed at the surfaces of cathodes,[159] detected in certain experiments by its photo-electric action and suggested by J. J. Thomson as producing the electron emission caused by positive ion impact.[160] Thus we should have, for the heat of neutralization,

$$\phi_+ = (1-r)V_i - \phi_- + L. \tag{52}$$

[158] Schottky and von Issendorff, Zeits. f. Physik **26,** 85 (1924); Compton and Van Voorhis, Proc. Nat. Acad. Sc. **13,** 336 (1927); Van Voorhis, Phys. Rev. **30,** 318 (1927).

[159] Geiger u. Scheel, "Handbuch der Physik," XIV, pp. 173, 183–188, 234.

[160] Thompson Phil. Mag. **48,** 1 (1924); **2,** 675 (1926).

The fraction (r) would be expected to be >0.5, from considerations of reflecting power and subtended solid angle of the metal surface. Compton and Van Voorhis' conclusion[158] that ϕ_+ is very nearly zero must now be retracted in the light of further evidence which has proved that their original results were affected by an unsuspected new complication now to be described. All that can at present be said is that ϕ_+ must be given by Eq. (52), with $1>r>0$.

A second factor which complicates and may considerably reduce the heating of electrodes by positive ions, and which appears to have been neglected in previous discussions of this question, is the *retention by the neutralized ion (atom) of a portion of its incident kinetic energy*. It has hitherto been assumed that the neutralized particle, if it was not permanently adsorbed on the electrode, departed from it with negligible energy (as, for example, in thermal equilibrium with it). However, this is not the typical behavior of neutral molecules incident upon a surface of different temperature, for it is found that such molecules only partially come into thermal equilibrium with the surface (except in the case of adsorption followed by evaporation, which is probably not the behavior of neutralized ions since these in general have kinetic energies far in excess of the heat of adsorption). The degree to which a reflected molecule adjusts its energy to that of the reflecting surface is described by its *accomodation coefficient*, defined by

$$a=(E_i-E_r)/(E_i-E_s) \tag{53}$$

where E_i, E_r and E_s are, respectively, the kinetic energy of incidence, that of reflection and that characteristic of the temperature of the reflecting surface. Thus complete adjustment at impact, as in case of adsorption and reevaporation, is described by $a=1$, while specular reflection corresponds to $a=0$. Some values of accomodation coefficient are given in Table X.

TABLE X. *Accomodation coefficients a.*

Gas	Metal	Metal Temperature	a	Ref.
H_2	Pt	20°C	0.26	161
CO_2	Pt	"	0.87	"
N_2	Pt	"	0.87	"
H_2	Pt	−100	0.25	162
H_2	Pt	+200	0.15	"
He	Pt	−100	0.49	"
He	Pt	+200	0.37	"
H_2	W	+1500	0.19	163
N_2	W	+1500	0.60	"
H_2	Pt	~20	0.278	164
O_2	Pt	"	0.800	"
CO_2	Pt	"	0.807	"
He	Pt	"	0.338	"
A	Pt	"	0.857	"
Ne	Pt	"	0.653	"

[161] Knudsen, Ann. d. Physik **34**, 593 (1911).
[162] Soddy and Berry, Proc. Roy. Soc. **84**, 576 (1911).
[163] Langmuir, J. Am. Chem. Soc. **37**, 425 (1915).
[164] Knudsen, Ann. d. Physik **46**, 641 (1915).

Compton and Van Voorhis, in some unpublished work still in progress,[165] have shown that the heating of a cool tungsten cathode by 50 volt positive ions of A, Ne and He is of the order of only 0.8, 0.65 and 0.43, respectively, of the expected watts input given by current to cathode times accelerating voltage drop. This is under conditions in which there is no possibility of energy dissipation by collisions of the incoming ions with gas molecules. It also neglects any contribution of the heat of neutralization ϕ_+ whose inclusion would tend further to diminish the above fractions. These fractions are strikingly similar to the corresponding accomodation coefficients, which is probably a coincidence, since the atom energies in this case are about a thousand times those of the atoms referred to in Table X and it would be quite surprising if a did not vary considerably over this range. The unexpectedly low heating is due to two factors (1) this effect of an accomodation coefficient less than unity and (2) the fact that part of the current is not incoming ion current, but outgoing electron current arising from action of excited atoms on the electrode,—as has been proven by supplementary tests.[142] Further possible evidence of this retention of energy by the neutralized particles is cited in the following section.

Thus far no quantitative estimates of the accomodation coefficient appropriate to impact of ions against electrodes are available. Until these have been determined, all estimates of the heating of cathodes by ion bombardment must be held in abeyance. However, the equation for this heating per positive ion of initial energy $\bar{V}$ may be stated as

$$H_+ = a\overline{V}_+ + \phi_+ = a\,\overline{V}_+ + (1-r)V_i - \phi_- + L. \tag{54}$$

Momentum transfers at electrodes due to impact of charged particles against electrodes, do not cause a pressure against the electrodes if, as is usual, these particles acquired their momenta entirely as the result of attraction by these electrodes, since the momentum gained by the electrode by attraction is exactly neutralized by the impact. Pressure against an electrode may, however, arise from impact of particles whose speeds were acquired in a field between other electrodes, or from the reaction of spontaneously emitted particles such as thermoelectrons from a hot cathode[166] or neutral atoms from a volatilizing electrode.[167]

Furthermore, if ions which are drawn to an electrode by an adjacent field, retain after neutralization a portion of their incident kinetic energy, as in case of an accomodation coefficient less than unity, the momentum of escape would be imparted to the electrode producing a pressure upon it. This would offer a reasonable explanation of the rather large pressure against the cathode of a copper vacuum arc, which led Tanberg[168] to the rather unprepossessing

[165] Compton and Van Voorhis, to be reported at the April meeting of the American Physical Society.

[166] Duffield, Burnham and Davis, Proc. Roy. Soc. Lon. **97**, 326 (1920).

[167] Beer and Tyndall, Phil. Mag. **42**, 956 (1921); Tyndall, *ibid.* **42**, 972 (1921); Sellerio, *ibid.* **44**, 765 (1922).

[168] Tanberg, Phys. Rev. **35**, 294 (1930).

conclusion that the copper atoms which evaporate from the surface, have kinetic energies characteristic of about 500,000°C. For example, the observed order of magnitude of this pressure would follow if half the current at the cathode is carried by positive ions, and if these retain on the average one-tenth of their kinetic energy after neutralization.

Langmuir[169] has discussed the related problem of momentum imparted to the gas, and the resulting pressure in discharge tubes.

"*Sputtering,*" *or distintegration of an electrode subjected to positive ion bombardment* is a well known and often troublesome phenomenon. Although discovered[170] as early as 1852 its behavior and explanation both remain rather obscure.

Most investigations have dealt with the rate of loss of weight of the cathode in a glow discharge, as a function of the nature of the metal, nature of the gas, cathode fall, current density, gas pressure, cathode temperature or geometry of tube. The most striking observation is the great difference observed with different metals and gases. Some of these observations are shown in Table XI.

TABLE XI.

Observer	Gas	Rate of sputtering in descending order
Crookes[171]	air	Pd, Au, Ag, Pb, Sn, Pt, Cu, Cd, Ni, Ir, Fe, Al, Mg
Kohlschütter[172]	N_2	Ag, Au, Pt, Pd, Cu, Ni
Blechschmidt[173]	A	Cd, Ag, Pb, Au, Sb, Sn, Bi, Cu, Pt, Ni, Fe, W, Zn, Si, Al, Mg
Güntherschulze[174]	H_2*	Bi 1470, Te 1200, As 1100, Tl 1080, Sb 890, Ag 740, Au 460, Pb 400, Zn 340, Cu 300, C 262, Sn 196, Fe 68, Ni 65, W 57, Co 56, Mo 56, Mn 38, Cd 32, Al 29, Cr 27, Ta 16, Mg 9
"	O_2*	Zn 1030, Tl 650, Ag 614, Au 423, Pb 320, Cu 236, Sn 227, Fe 86, Mo 80, W 49, Ni 52, Cd 28

* Note: Numbers give rate of sputtering in Mg/amp. hr. under the conditions of cathode fall $V_c = 770$ volts and current density about 7 milliamps. per cm^2.

A relation which holds with fair accuracy with most gases gives the mass of cathode material sputtered in unit time[175]

$$m = K(V_c - V_0), \tag{55}$$

where K and V_0 are constants characteristic of the gas and metal and V_c is the cathode fall. Thus the rate of sputtering is about proportional to the excess

[169] Langmuir, Jour. Frank. Inst. **196,** 751 (1923).
[170] Grove, Phil. Trans. 1852.
[171] Crookes, Proc. Roy. Soc. Lon. **50,** 88 (1891).
[172] Kohlschütter, Zeits. f. Elektrochem. **15,** 316 (1909); Jahrbuch d. Radioakt. **9,** 355 (1912)
[173] Blechschmidt, Ann. d. Physik **81,** 999 (1926).
[174] Güntherschulze, Zeits. f. Physik **36,** 563 (1926); **38,** 575 (1926).
[175] Grandquist, Kgl. Akad. Stockholm, **54,** 595 (1897); Holborn and Austin, Wiss. Abhand. d. Phys. Tech. Reichsanstalt, **4,** 101 (1903).

of cathode fall of potential above V_0. V_0 usually lies between 350 and 550 volts and is generally about 450 volts. The constant K is roughly proportional to the fourth root of the atomic weight of the gas,[176] unless the rate of sputtering is accelerated by chemical action.[177] There have been suggestions[172] also that $K = K_0 A/n$, where A is the atomic weight of the metal, n is an integer between 1 and 4 which is in some cases (but not always) equal to the valency, and K_0 is a constant characteristic of the gas. The value of K_0 for N_2 is about 0.0004.

The approximate validity of Eq. (55) does not necessarily imply that the rate of sputtering is causally related to the cathode fall V_c, since, in order to vary the cathode fall, the gas pressure or current density were varied. Güntherschulze[177] found that the rate of disintegration of the cathode depends on the distance of the anode, and also on the gas pressure and geometry of the cathode, in such a way as to suggest that the cathode loses weight by the process of diffusion of metal atoms away from the region of their maximum partial pressure just outside the cathode surface toward the anode where their partial pressure is zero. According to him, at constant current density

$$m = CV_c/pD \tag{56}$$

where C is a constant characteristic of the gas, p is the gas pressure and D is the distance to the anode.

All the above relations and experiments, while useful empirically to describe the rate of cathode disintegration in ordinary glow discharges, are ill adapted to yield information regarding its fundamental cause, on account of the complexity of the conditions. Much more incisive are experiments which give the rate of distintegration caused by bombardment of ions of known energy under conditions in which collisions of sputtered atoms with gas molecules do not occur to retard their escape. Such experiments have been performed by The General Electric Co., Ltd.[178] and by Kingdon and Langmuir.[179] Here electrons from a hot filament were used to ionize the gas whose pressure was too low appreciably to interfere with the escape of the atoms sputtered from the cathode. The positive ions were drawn to the cathode with known and fairly homogeneous velocities. The results again led to a relation $m = (V - V_0)$ like Eq. (55) but with much lower values of V_0, indicating that in the case of the glow discharge the positive ions striking the cathode have lost by collisions with gas molecules most (at least three-fourths) of the energy gained in the cathode fall.

The General Electric Co., Ltd. found no case in which V_0 exceeded 100 volts. The effectiveness of gas ions in disintegrating a tungsten cathode increase in the order H_2, He, N_2, Ne–He, Hg, A. Kingdon and Langmuir tried the effect of ion bombardment in disintegraing the monatomic layer of thor-

[176] Kohlschütter and Müller, Zeits. f. Elektrochem. **12,** 365 (1906); Kohlschütter and Goldschmidt, *ibid.* **14,** 221 (1908); Blechschmidt, Ann. d. Physik. **81,** 999 (1926).

[177] Güntherschulze, Zeits. f. Physik. **38,** 575 (1926).

[178] General Electric Co., Ltd. Phil. Mag. **45,** 98 (1923).

[179] Kingdon and Langmuir, Phys. Rev. **22,** 148 and 357 (1923).

ium on tungsten and found, for 150 volt impacts, that one thorium atom was knocked off for every 700,000 impacting ions of He, 45 of Ne, 23 of Hg, 12 of A, 12 of Cs, and that hydrogen ions were entirely ineffective. The very small sputtering power of hydrogen and helium is believed to be due to the relatively great penetrating power of these atoms, so that their energy is dissipated too deep within the metal to cause surface disintegration. Fig. 11 gives some illustrative data.

Holst[180] has observed sputtering of tungsten in argon rectifiers at voltages as low as 25 volts.

Oliphant[141] gives reasons for suspecting that it is not alone the kinetic energy of the impacting ion which is responsible for sputtering, but that the fact that the ion is an electric charge is also a factor.

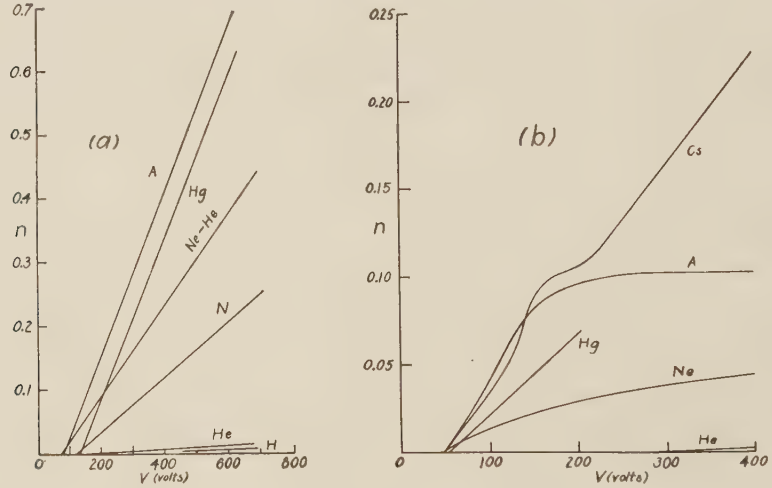


Fig. 11. (a) Sputtering of W (General Electric Co., Ltd.); (b) sputtering of surface Th atoms from a W filament (Kingdon and Langmuir). n = number of sputtered atoms per impinging positive ion.

There is spectroscopic and magnetic deflection evidence that sputtering may consist in the expulsion of neutral atoms, whose velocities are of the order of those occurring in evaporation processes,[181] although these may acquire charges, coalesce into aggregates, etc., in their subsequent history. Tungsten, molybdenum and carbon sputtered from filaments of these materials in argon are capable of acquiring either positive or negative charges. This results in a very interesting type of discharge, the streamer discharge, which may occur when these materials are sputtered in arcs in argon (also in neon).[182]

When thermionic currents of 50 to 100 volts are passed through argon at pressures of only a few microns, a tungsten filament sputters rapidly and the

[180] Holst, Physica **4**, 68 (1924). See also A. W. Hull, Trans. Amer. Inst. E.E., **49**, 753 (1923).

[181] von Hippel, Ann. d. Physik **80**, 672 (1926); Baum, Zeits. f. Physik **40**, 686 (1927).

[182] I. Langmuir, Science **60**, 392 (1924)

sputtered material deposits on the bulb particularly behind the anode, indicating that it becomes negatively charged.[183]

There are two principal theories of sputtering, the theory of local high temperatures and the theory of direct momentum transfers. The theory of local high temperatures has been developed most extensively by von Hippel[184] on the idea that the energy of the impinging ion raises an area around the point of impact to a temperature which is instantaneously very high, but which sinks rapidly as the energy spreads out and is dissipated by the thermal conductivity of the metal. He finds

$$m \propto \frac{i}{e(1+C_e)T} \epsilon^{-q/RT} \Delta F \Delta t,$$

where i is the current density, e is the electronic charge, C_e is the number of secondary electrons emitted per impinging positive ion, T is the average temperature of the average area ΔF from which evaporation occurs during a time Δt, R is the gas constant and q is the heat of evaporation of the metal. ΔF, T and Δt may be estimated from considerations of the kinetic energy of the ion and the thermal conductivity of the metal as shown in the second reference (184). While this equation gives the right general sort of variation among the parameters, it rests on necessarily very crude approximations. An apparently serious objection to any such high temperature evaporation theory lies in the fact that it would predict, in the case of a metal like tungsten, an enormous secondary (thermionic) electron emission in proportion to the amount of sputtering, whereas this is not found to be true. It is possible, however, that such emission is limited to a small value by space charge on account of the extremely small dimensions ΔF of the effective emitting areas.

Kingdon and Langmuir[179] showed that the rate of sputtering of thorium from a tungsten surface can be quite accurately explained, and the minimum sputtering voltage (about 50 volts) roughly calculated on the following assumptions: As the first step in the process, a surface atom of thorium is struck by an ion and driven into the underlying tungsten, forming a depression. When this depressed thorium atom is struck by a second ion, the ion is elastically reflected and, on its way back, may strike one of the surrounding thorium surface atoms and dislodge it, provided the energy which it can communicate to the atom, according to the laws of momentum transfer, exceeds the atomic heat of evaporation. The total chance of removing a thorium atom as the result of N ion impacts against any particular atom is given by

$$(1-\theta) = \mu(3P_2+5P_3+6P_4+6P_5), \tag{57}$$

where

$$P_n = \frac{1}{n!}(Np)^n \epsilon^{-Np}$$

[183] I. Langmuir, Phys. Rev. **2**, 475 (1913).

[184] von Hippel, Ann. d. Physik **81**, 1043 (1926); Blechschmidt and von Hippel, *ibid.* **86**, 1006 (1928).

is the probability that the same atom will be struck n times while N ions strike the surface, in terms of the probability p that a given ion strikes a given atom. θ is the fraction of the surface covered by thorium atoms and μ is the probability that an ion which has been reflected from a depressed thorium atom will remove one of the neighboring atoms on its rebound. Similarly, the minimum voltage $\bar{V}_s$ for sputtering is given in terms of the atomic heat of evaporation of thorium E_v from tungsten (1.40 $(10)^{-11}$ ergs/atom) by

$$V_s = \frac{300E_v}{4em_g m_c}\left[\frac{(m_g+m_c)^2}{m_g-m_c}\right]^2 \tag{58}$$

where m_g and m_c are the masses of the impacting ion and the target (cathode) atom, respectively.

Equation (57) accurately fits the experimental data when reasonable values of the parameters p and μ are chosen. Equation (58) gives about the right value of V_s for Ne and A, but too high values for H, He, Cs and Hg ions. The evidence cited above indicates that the discrepancy with H and He is probably due to their great penetrating power, while the discrepancy with the heavy ions of Hg and Cs may well be due to their driving the thorium atom so far into the tungsten at the first impact that the reaction to the second impact is by the entire tungsten wall rather than by the thorium atom alone, as assumed in Eq. (58).

Holst's observation of sputtering of tungsten in argon rectifiers at voltages as low as 25 volts led him to postulate a mechanism different from that assumed to explain the sputtering away surface thorium atoms. He suggests that it depends upon the transfer of the requisite amount of energy to a tungsten atom at a single direct impact, according to which

$$V_s = \frac{300E_v}{4e}\,\frac{(m_g+m_c)^2}{m_g m_c}\,. \tag{59}$$

This expression gives very small values for V_s, (*e.g.*, W in A, 16; W in Ne, 27; Pt in A, 10; Pt in Ne, 17; Cu in A, 3.3; Cu in Ne, 4.2). The accuracy of these values is very uncertain.

These remarks will serve to emphasize the fact that cathode sputtering, though an important phenomenon of gas discharge, is still very inadequately understood. It may be significant that the ion bombardment fractures the surface crystalline layer. (See micrographs of Baum[181] and evidence from electron emission by Reynolds).[105] If sub-microscopic dust were thus produced, it would be quickly evaporated when it passed into the discharge by the heat of recombination of ions and electrons on its surface. In the case of tungsten the evaporation is very slow and such incandescent centers of recombination can actually be seen (and investigated optically by a Tyndall beam) when tungsten is sputtered in argon.[182] In the case of the more volatile metals, such a process might possibly explain the atoms of sputtered material which have been taken as proving the evaporation theory of sputtering.

(2) **Disappearance of electrons and ions by recombination.** That electrons and ions may recombine has been known ever since their existence in a conducting gas was recognized. In fact all emission of light from glowing gases was at one time ascribed to this process. Yet attempts to produce light by mixing together ions and electrons under controlled conditions have proved singularly fruitless. Within the last ten years, however, certain restrictions on recombination have become apparent, and their recognition has greatly simplified the interpretation of the hitherto seemingly conflicting observations. These restrictions, whose realization is due principally to Bonhoeffer and Franck, are based upon the following idea.

Free electrons in an ionized gas possess translational kinetic energies anywhere from the value characteristic of the gas temperature upward, depending upon the magnitude of the electric field and gas pressure and the initial conditions of their formation. Whenever one of these electrons comes within the range of attraction of a positive ion, it would be expected, on classical theory, to move past the ion in an hyperbolic orbit. But permanent attachment (recombination) involves settling down into an elliptic orbit, which can only occur if the electron loses a part of its energy. The only ways of losing energy are by collision with a third body during the interval of close approach, or by radiation. The first of these involves a three body collision which is extremely improbable except under very special conditions which will be discussed below. The second involves the radiation of energy outside of the regular line spectrum, and the observed weakness of such radiation proves that recombination in this manner also is highly improbable. Thus we conclude that recombination of electrons and ions is a far less probable occurrence than was formerly supposed, when it was believed that every approach of an electron to an ion resulted in its capture. We shall proceed to a closer investigation of recombination under these restricted conditions.

Recombination on surfaces is a three body process in which the surface acts as the third body to absorb excess energy and momentum. The electrons and ions do not need to arrive simultaneously at a point on the surface, but one may arrive first and remain on the surface as a surface charge until the other arrives to combine with it. Herein is the explanation of the effectiveness of surfaces in causing disappearance of ions by recombination. Dust particles, submicroscopic particles sputtered from electrodes, insulating surfaces or floating electrodes all act in this way. As will be pointed out in detail later, insulated surfaces are charged with respect to the surrounding gas to a positive or negative potential, depending on whether ions or electrons carry the maximum random current density. Except in very special cases near cathodes, the random current of the electrons is always in excess owing to their small mass and resulting rapid motion. Thus isolated surfaces in general take up negative potentials with respect to the surrounding gas to such a degree that the electrons are repelled and the positive ions are attracted sufficiently to cause electrons and ions to arrive at equal rates and combine at the surface. Such surfaces in an intensely ionized gas may become highly heated, small floating metallic particles being often evapo-

rated or heated to incandescence. The energy comes from the heat of combination, the initial kinetic energies of the combining charges, and the energy gained by the positive ions in falling through the difference of potential between the surface and the surrounding space.

Recombination in the gas may occur, in a manner analogous to that described in the case of surfaces, by two *successive* impacts involving three bodies, (as distinguished from the simultaneous impact of three bodies), or it may occur by the direct union of an electron and ion without interaction of a third body. The first of these processes occurs in the following way: Electron attachment, *i. e.*, the formation of negative ions may occur at the collision of an electron with a neutral molecule. Then, if this negative ion collides with a positive ion, the electron transfers its attachment to the ion, releasing the neutral molecule. The neutral molecule may carry away the excess energy, if any, but we shall see that there often is no excess energy in such a case. The rapidity of this recombination process is limited, of course, by the rapidity of the slower one of the two cooperating processes (1) electron attachment to form negative ions, (2) combination of negative and positive ions.

Electron attachment to form negative ions has been studied most extensively by Loeb and his pupils.[185] The probability of attachment at an impact differs so enormously in different gases and is relatively so large in some gases and vapors which are very difficult to eliminate as impurities, that most of the conclusions are qualitative rather than quantitative. N is the total number of collisions made per second by an electron in the gas at atmospheric pressure and ordinary room temperature, assuming the electron mean free path to be the classical value of $4\sqrt{2}$ times that of a gas molecule. The attachment "constant," is not really a constant, since it is somewhat dependent upon the electron velocities. Furthermore it is not always an additive property in gaseous mixtures, since water vapor appears to cause more attachment when present as an impurity than would be expected from the value of n given in Table XII. n does not seem to be related with any obvious property of the molecules, unless it be roughly with their electronegative character. A prevalent idea that the molecular dipole moment is the controlling factor is obviously untenable, NH_3 and H_2O possessing strong moments and N_2 and Cl_2 no moments, to take extreme examples.

Several interesting deductions may be drawn from Table XII. In pure gas at atmospheric pressure an electron would remain free on the average for only about 7 $(10)^{-4}$ sec. in CO, about 4.7 $(10)^{-9}$ sec. in Cl_2, unless strong fields should so speed up the electrons as to render attachment still more improbable. Obviously most experimental work which has been done on ion mobilities and all the early work on recombination have been done

[185] Loeb, Kinetic Theory of Gases, pp. 507–515; J. Frank. Inst. **197**, 45 (1924); Wahlin, Phys. Rev. **19**, 173 (1922); Cravath, *ibid.* **33**, 605 (1929).

under conditions in which ions, rather than electrons, were the negative carriers of charge. In "vacuum tube" discharges, on the other hand, with pressures of the order of 1 mm or less, the average free life of electrons in many gases is sufficiently long to insure that most if not all the phenomena of negative charge are due to free electrons, even neglecting the increased

TABLE XII. *Attachment constant n (average number of electron impacts which result in one attachment to form a negative ion), compared with number N of electron impacts per second against gas molecules.*

Gas	n	N
Noble gases, N_2, H_2		
CO	$1.6\ (10)^8$	$2.22\ (10)^{11}$
NH_3	$9.9\ (10)^7$	2.95
C_2H_4	$4.7\ (10)^7$	3.75
C_2H_2	$7.8\ (10)^6$	4.12
C_2H_6	$2.5\ (10)^6$	4.85
N_2O	$6.1\ (10)^5$	3.36
C_2H_5Cl	$3.7\ (10)^5$	5.45
Air	$2.0\ (10)^5$	2.17
O_2, H_2O	$\sim 4.0\ (10)^4$	2.06, 2.83
Cl_2	$<2.1\ (10)^3$	4.50

effectiveness at low pressures of the electric field in maintaining electrons in the free state. (This effectiveness is a function of E/p.) Also the tremendous effect of some impurities on electron attachment is obvious. For example the rate of attachment CO would be increased *at least* 100 percent by the presence of 0.025 percent of oxygen or water vapor.

Combination of negative and positive ions is not restricted, since the neutral molecule to which the electron is attached may function as the third body to carry off energy in excess of that corresponding to some excited state of the combining positive ion and electron. There is no direct evidence regarding the probability of such an energy transfer, and therefore no knowledge regarding the probability of combination of two oppositely charged ions which collide with a relative velocity greater than that which they would acquire in falling together from infinity. The two principal theories of ion combination neglect this probability, and assume that the combining ions have lost a part of the energy, gained by their mutual attraction, at collisions with neutral molecules, so that once they come within each other's sphere of influence, they remain there until recombination is effected.

The rate of ion recombination is jointly proportional to the concentrations n_1 and n_2 of the positive and negative ions, and is therefore

$$-\frac{dn_1}{dt} = -\frac{dn_2}{dt} = \alpha n_1 n_2, \tag{60}$$

where α is the "*coefficient of recombination.*" Table XIII gives some typical values:

TABLE XIII. *Coefficient of recombination α at atmospheric pressure and ordinary temperature in units of $cm^3\ sec^{-1}$. (Ionization produced by X-rays).*[186]

Gas	Air	CO_2	O_2	H_2	SO_2	N_2O	CO
$(10)^6$	1.71	1.67	1.61	1.44	1.43	1.42	0.85

From Table XIII we see that, if there were n positive and n negative ions per cm^3 in air, the rate of recombination would be $1.71\ (10)^{-6}\ n^2$.

The number of collisions between gas molecules of types 1 and 2 is given by kinetic theory as[187]

$$\text{no. collisions} = 2(2\pi/3)^{1/2} n_1 n_2 \sigma_{12}^2 (C_1^2 + C_2^2)^{1/2}\ \text{cm}^{-3}\ \text{sec}^{-1} \qquad (61)$$

where C_1 and C_2 are their root mean square velocities and σ_{12} is the sum of their effective collision radii. From this we find that in air at room temperature, the number of collisions per cm^3 sec. between 2 n molecules is $1.6\ (10)^{-10} n^2$. Comparing this with the value $1.71\ (10)^{-6} n^2$ above, we see that the *actual rate of recombination is about 10,000 times greater than it would be if due simply to ordinary kinetic theory collisions of positive with negative ions.* It is therefore concluded that mutual electrostatic attraction greatly accelerates the process of recombination. This effect of attraction has been differently handled by Langevin and Thomson.

Langevin's theory of recombination of ions[188] is based on the assumption that neighboring oppositely charged ions are drawn toward each other with velocities determined by the attractive force and by their mobilities* μ_1 and μ_2. When their distance apart is r, the attractive field is e/r^2 and the relative velocity of approach is $e(\mu_1+\mu_2)/r^2$, provided there are no other ions within a distance comparable with r. If there are n_2 negative ions per cm^3, the rate at which negative ions will cross the surface of a sphere of radius r about a positive ion is

$$4\pi r^2 \cdot n_2 \cdot e(\mu_1+\mu_2)/r^2 = 4\pi e n_2 (\mu_1+\mu_2),$$

which is independent of r. Thus the average time required for a negative ion to reach this particular positive ion is $1/4\pi\ en_2\ (\mu_1+\mu_2)$ sec. But since there are n_1 positive ions, the average time interval between combinations

[186] Calculated from values collected by Thomson, "Conduction of Electricity Through Gases," 3 ed., p. 40.

[187] See Jeans, "Dynamical Theory of Gases," 2 ed., p. 275 and also Eq. (83) in Section (D 1 b) of this article. Thomson, "Cond. Elec. Through Gases," 3 ed., p. 43 uses an equation whose coefficient is 2.45 times too large.

[188] Langevin, Ann. de Chem. et de Phys. **28,** 287, 433 (1903); Townsend, "Electricity in Gases," pp. 206–220, in which Langevin's theory is extended and certain limits of applicability are discussed.

* The "mobility" is the constant μ in the relation $\bar{v} = \mu E$ between the field strength E and the average rate of drift $\bar{v}$ in the direction of the field. As pointed out below, μ is not in general a constant, but is itself a function of E. At high gas pressures and weak fields, μ is sensibly independent of E.

with *any* positive ion is $1/4\pi\ e\ n_1\ n_2\ (\mu_1+\mu_2)$. The reciprocal of this gives the rate of recombination, and therefore α, as

$$\alpha n_1 n_2 = 4\pi e(\mu_1+\mu_2)\cdot n_1 n_2. \tag{62}$$

This theory implicitly assumes that ions recombine whenever they are drawn together, which is self consistent since it is only justifiable to use mobilities when the moving ions lose energy so rapidly by collisions that they are always in or near a state of terminal speed and under such conditions the kinetic energy retained by the ions is insufficient to cause their separation once they have come close together. This consideration suggests that Langevin's theory should be most applicable at high gas pressures, and should fail at low pressures, as is the case. Also, since μ_1 and μ_2 vary inversely with pressure, α should also vary thus according to the theory. This is true at high pressures (such as several atmospheres) but not at low pressures. The absolute values of α calculated from Eq. (62) are approximately correct at high pressures, but are as much as 100 times too large at the lower pressures at which measurements of α and μ are made (such as 0.1 atmosphere).

Townsend[188] has extended Langevin's theory to include approximately the effect of thermal agitation, for so long as the ions have not approached so closely that the work to separate them greatly exceeds the energy of thermal agitation, it is always possible for the ions to escape combination. This fact, and the influence of neighboring ions, introduce corrections to Langevin's theory in the right direction. The most elegant treatment of thermal energies is, however, due to Thomson.

Thomson's theory of recombination of ions[189] is based on the idea that combination occurs if one of two neighboring ions of opposite sign collides with a neutral molecule and loses so much kinetic energy that it cannot escape against the attraction of the other. If v is the relative velocity of the two ions of masses M_1 and M_2 at a distance r apart, then they will execute closed orbits about each other and ultimately combine provided r is less than r_0 in the relation

$$\frac{1}{2}\frac{M_1 M_2}{M_1+M_2}v^2 = \frac{e^2}{r_0}.$$

Thomson now assumes that the mean velocities u_1 and u_2 immediately after one of the ions has collided with a molecule are those characteristic of thermal agitation. (This is certainly not true if the ions have been drawn close together, but it appears to be sufficiently justified at the critical distance r_0 which alone is important for the present purposes.) Then, since $v^2 = u_1^2 + u_2^2$,

$$\tfrac{1}{2}M_1 u_1^2 = \tfrac{1}{2}M_2 u_2^2 = \tfrac{3}{2}kT$$

$$v^2 = 3kT\left(\frac{M_1+M_2}{M_1 M_2}\right),$$

[189] Thomson, Phil. Mag. **47**, 337 (1924); "Conduction of Electricity Through Gases," 3 ed. pp. 44–57.

whence the critical distance r_0 is given by

$$\frac{e^2}{r_0}=\frac{3}{2}kT \tag{63}$$

and is seen to vary inversely as the absolute temperature T.

Now the chance that ion 1 will collide with a molecule while passing ion 2 within the critical distance r_0 is found in terms of its mean free path λ_1 to be

$$w_1=1+2\left[\frac{(\epsilon^{-2r_0/\lambda_1}-1)}{(2r_0/\lambda_1)}+\frac{\epsilon^{-2r_0/\lambda_1}}{(2r_0/\lambda_1)}\right]$$

and a corresponding expression holds for the chance w_2 that ion 2 will collide. Thus the total chance that one of the two ions which come within a distance r_0 will collide while within this distance is

$$w_1+w_2-w_1w_2,$$

the last term which represents the chance of *both* ions colliding being already included in w_1 and w_2.

Finally, the number of times that ions come within a distance r_0 of each other is approximately $\pi r_0^2 n_1 n_2 (u_1^2+u_2^2)^{1/2}\text{cm}^{-3}\ \text{sec}^{-1}$. Since every one of these that collides is supposed to enter into recombination with its neighboring ion, the rate of recombination is

$$\alpha n_1 n_2=\pi r_0^2(u_1^2+u_2^2)^{1/2}(w_1+w_2-w_1w_2)\cdot n_1n_2.$$

Substituting for r_0 from Eq. (63) we obtain

$$\alpha=\frac{4\pi e^4}{9k^2T^2}(u_1^2+u_2^2)^{1/2}(w_1+w_2-w_1w_2). \tag{64}$$

At high pressures w_1 and w_2 approach unity, so that α should increase with pressure, but approach a constant value at high pressures. Actually this is found to be true[190] except that at still higher pressures α is found again to diminish with increasing pressure. This marks the point at which Langevin's theory begins to describe the phenomena more accurately than Thomson's.

Quantitatively, Eq. (64) fits the facts throughout a considerable range of pressures as accurately as could be expected in view of the very considerable uncertainty regarding the ionic mean free path and the exact nature (mass) of the ions,[191] which may often be ions of impurities or of clusters.

In Langevin's theory the rate of recombination is limited by the rate at which ions are drawn together by their mutual attraction; in Thomson's theory it is limited by the rate of diffusion of ions into regions in which the attractive force can hold them together. Additional support of Thomson's

[190] Langevin, Comptes Rendus **137**, 177 (1902); McClung, Phil. Mag. **6**, 283 (1902).
[191] Loeb, Phys. Rev. **32**, 81 (1928).

theory as against Langevin's for pressures of the order of an atmosphere and downward has recently been deduced by Loeb and Marshall[192] from the observation[193] that recombination is abnormally rapid for a short time after the formation of the ions, especially if the ionizing agent is one, like α particles, which creates an uneven distribution of ions. This was correctly explained[194] as due to the fact that the actual effective initial ion concentrations are much greater than the value calculated on the basis of even distribution throughout the entire volume, and it was shown that the correct values of α are those taken after lapse of sufficient time to permit uniform distribution. Loeb and Marshall pointed out that the very fact that an even distribution is attained shows that diffusion is a more important element in the life of the ions than is inter-ion attraction. They further analyzed the problem by considering the ions as executing Brownian movements subject to the action of their electric fields and showed that, at atmospheric pressure and room temperature, their mutual fields were practically a negligible factor, as compared with diffusion, in bringing the ions together from distances beyond about one free path. At several atmospheres, however, the field is important.

Recombination of free electrons and positive ions[195] is that type of recombination which, next to recombination on surfaces, is of the greatest interest in gas discharges. Its existence cannot be assumed simply because of the production of light in gas discharges because it is readily shown that light is principally due, in most cases, to excitation of atoms rather than to recombination. For example, the positive column of a glow discharge emits more intense light than does the negative glow. Yet measurements show[196] that electron and ion concentrations are each of the order of 100 times larger in the negative glow than in the positive column, and their velocities are lower. If light in both regions were due to recombination, it should be 100^2 times more intense in the negative glow because of greater concentrations, and still more than this because of the lower velocities which favor recombination. Thus, certainly, recombination plays a negligible role in the production of light in the positive column. In further support of this conclusion we shall see in Part II of this article that the radial concentration and potential gradients as well as the relation between axial potential gradients and tube diameters in cylindrical positive columns are such as to prove that all other types of recombination in the positive column with low pressures of gas are negligible in comparison with recombination on the tube walls.

This does not prove, however, that the negative glow light may not be due appreciably or perhaps mostly to recombination. That recombination is important here is proved by the fact that spectrum lines from the negative

[192] Loeb and Marshall, J. Frank. Inst. **208,** 371 (1929).

[193] Plimpton, Phil. Mag. **25,** 65 (1913); Rümelin, Ann. d. Physik **45,** 821 (1914).

[194] Plimpton, Phil. Mag. **25,** 65 (1913); Townsend, "Electricity in Gases," p. 200.

[195] For an excellent review of this subject see Seeliger, Phys. Zeits. **30,** 329 (1929); also a discussion of some calculated orders of magnitude by Atkinson, Zeits. f. Physik **51,** 188 (1928).

[196] Compton, Turner and McCurdy, Phys. Rev. **24,** 597 (1924).

glow show a Doppler widening[197] of the magnitude to be expected if the radiating atoms are moving with velocities of the same order as positive ion velocities,—distinctly more rapidly than molecular velocities,—whereas lines from the positive column have widths appropriate for emitters with ordinary thermal velocities. In fact, from Miss Dewey's experiments one would infer that, at least in her apparatus, practically all the negative glow light is due to recombination.

There are two additional criteria by which a recombination spectrum may be recognized. One of these, used effectively by Miss Hayner[198] and Kenty[199] is the relatively large intensity of the higher members of spectral series, as compared with their intensities in spectra due to excitation. This is explained by the fact that in combining, every electron has a chance to make transitions between the higher energy states as well as the lower ones, whereas in excitation most of the atoms are excited only as far as the lower states,—a fact proven by direct observation of intensities of excitation spectra under conditions of such low ion concentration as to make recombination completely negligible.

Kenty's observations with argon are particularly illuminating. He ionized the gas intensely by a low voltage arc from a hot cathode, then cut off the voltage (or else reduced it to a second value much too low to maintain the arc) and during this second period simultaneously photographed the spectrum and measured the velocity distribution of the electrons by Langmuir's probe method. A rotating sector disk and commutator enabled these intervals to be rapidly repeated and continuous measurements to be made. The intensity of high series members was a measure of the rate of recombination, and this was found to be greatly reduced as the mean velocities of the electrons was increased by means of the post-arc voltage. With no post-arc voltage the mean electron energy was 0.4 volt, the electron and ion concentrations were of the order of 10^{12} per cm^3, and the coefficient of recombination α was found to be about $2\ (10)^{-10}$. The number of kinetic theory collisions between positive ions and electrons may be calculated by Eq. (61), assuming the effective collision radius to be that of an argon atom and it is found that the actual rate of combination is 1/375 of the rate at which electrons and ions would be thus calculated to collide, which may be taken as a very rough upper limit to the *probability of combination* of argon ions with 0.4 volt electrons. We saw that positive and negative *ions* may combine 10,000 times faster than the rate of kinetic theory collisions, whereas positive ions and electrons combine several hundred times slower than the kinetic theory collision rate, even under the exceptionally favorable condition of unusually low velocity electrons.

Table XIV gives an example of the relatively greater intensity of higher series lines in the recombination spectrum.

[197] Dewey, Phys. Rev. **32,** 918 (1928).

[198] Hayner, Zeits. F. Physik **35,** 365 (1926).

[199] Kenty, Phys. Rev. **32**, 624 (1928).

TABLE XIV. *Intensities of argon lines $2p_{10}-md$ (Kenty).*[199]

m	Intensity	
	Recomb. Spectrum	Arc Spectrum
5	15	20
6	14	8
7	10	1
8	8	0
9	2	—
10	3	—

The other criterion of recombination is the existence of continuous bands stretching to the short wave-length side of series limits. Fig. 12 shows results of a photometric measurement of the intensity distribution in these bands in the case of caesium vapor[200] in the neighborhood of the absorption limits 3 $D_{2,3}$ and 2 $P_{1,2}$ and 1 S_0.

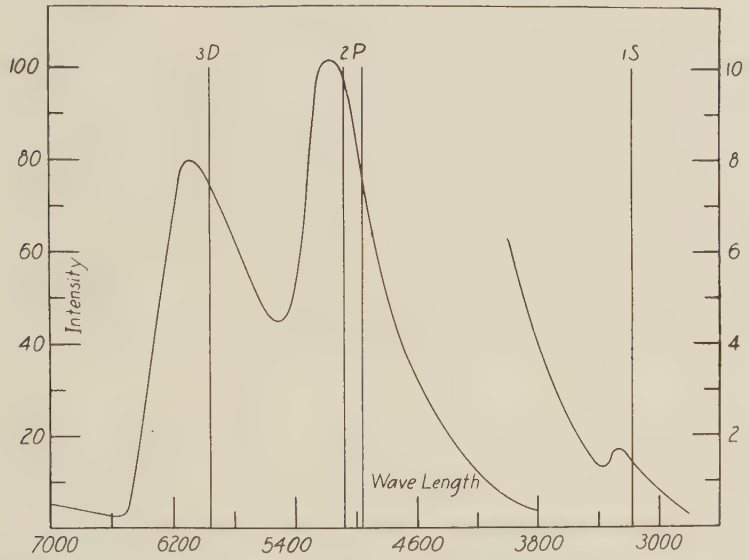


Fig. 12. Intensities of continuous recombination spectra beyond absorption limits in Cs vapor.

The interpretation of these bands is simple: The frequency associated with a state of energy $h\nu_i$ of an atom is the high frequency limit of the spectral series due to the falling of electrons into state i from states of higher energy. If an electron, whose kinetic energy was zero at a great distance, were to combine with an ion and fall directly into the energy state i, the emitted light would have the frequency ν_i of this series limit. But, actually, the electrons which combine have energies $\frac{1}{2}m\ v_2$ greater than zero. On classical theory these could not combine since they would execute hyperbolic orbits about the ion. But on quantum theory there is a certain probability,

[200] Mohler, Phys. Rev. **31**, 187 (1928); see especially Phys. Rev. Supp. **1**, 216 (1929).

which is a function of v, that they will combine. An electron of velocity v which combines causes emission of light of frequency ν given by

$$h\nu = h\nu_i + \tfrac{1}{2}mv^2. \tag{65}$$

Since the velocities v are continuously distributed (frequently with a Maxwellian distribution) it is obvious that ν will have all values from ν_i upward,[201] giving the continuous band.

The intensity distribution in the band depends obviously on two factors, (1) the distribution of electron velocities and (2) the probability of combination, or recombination coefficient characteristic of each velocity. The former factor $N(v)$ may be measured by Langmuir's probe method, so that the band intensity measurements may be used to estimate the coefficient of recombination $\alpha(v\,\nu_i)$ for electrons of any given speed v to form atoms in the energy state i. The equation for this may be written*

$$I(\nu)d\nu = h\nu \cdot \alpha(v\nu_i) N^+ N^-(v) dv \tag{66}$$

where $h\,d\nu = m\,v\,dv$ and $I(\nu)\,d\nu$ is the energy of frequency between ν and $\nu + d\nu$ radiated per second per unit volume and $N^-(v)\,dv$ is the concentration of electrons with velocities between v and $v+dv$. The recombination coefficient $\alpha(v\,\nu_i)$ is frequently written $v\,q(v\,\nu_i)$, where $q(v\,\nu_i)$ is the effective cross section for recombination of electrons of velocity v going into the i electron energy state of the atom. The total recombination coefficient $\alpha(v)$ for electrons of velocity v is, of course, $\alpha(v) = \sum(v\,\nu_i)$, and the ordinary total recombination coefficient is $\alpha = \int_0^\infty \alpha(v)\,dv$.

Mohler[202] has made a careful study of recombination spectra in caesium and helium by this method, and has shown that the coefficient of recombination $\alpha(v\,\nu_i)$ varies with electron velocities according to an equation of the form

$$\alpha(v\nu_i) = \frac{K}{\nu(\nu-\nu_i)^{1/2}} \text{ or } \frac{K}{\nu^2(\nu-\nu_i)^{1/2}} \tag{67}$$

and depends so much more on the term $(\nu-\nu_i)^{1/2}$ than on the term ν^n that it is impossible to say with certainty whether $n=1$ or 2, although the agreement is slightly better with $n=2$. Thus the recombination coefficient approaches infinity very close to a series limit, but diminishes rapidly at higher frequencies. In Eqs. 66, 67, ν and v are, of course, related by Eq. (65).

According to a general relation based on the principle of detailed statistical balancing between the probability of recombination $\alpha(v\,\nu_i)$ and the probability of absorption $B(\nu\,\nu_i)$ at a frequency ν corresponding to v by Eq.

[201] The fact that the bands are observed to extend a slight distance to the long wave length side of ν_i is probably due to a blurring of the limits by the Stark effect in highly ionized gases. (Robertson and Dewey, Phys. Rev. **31**, 973 (1928)).

* We here assume, for simplification, that continuous spectra associated with different limits do not overlap. If they do overlap, Eq. (66) may be expressed as a summation, or the experimental contributions from the different bands may be separated by extrapolation.

[202] Mohler, Phys. Rev. **31**, 187 (1928); Research Paper No. 46, U. S. Bureau of Standards, March 1929.

(65), Milne[203] has derived a relation $\alpha(v\nu_i) = K'B(\nu\ \nu_i)\nu^3/(\nu-\nu_i)^{1/2}$. Also it is well known[204] that $B(\nu\ \nu_i)$ is very closely proportional to $1/\nu^4$, at least in the X-ray region well beyond an absorption edge. Hence we should expect approximately $\alpha(v\ \nu_i) = K/[\nu(\nu-\nu_i)^{1/2}.]$ which is in quite satisfactory agreement with Eq. (67). It is this quantity $B(\nu\ \nu_i)$ which expresses the probability of photo-ionization. (See section A(2) .) The experiments on recombination are in rather better accord with the theory than are those on photo-ionization.[205]

It seems likely that the above theory of recombination is on a fairly satisfactory basis in so far that it enables recombination coefficients to be

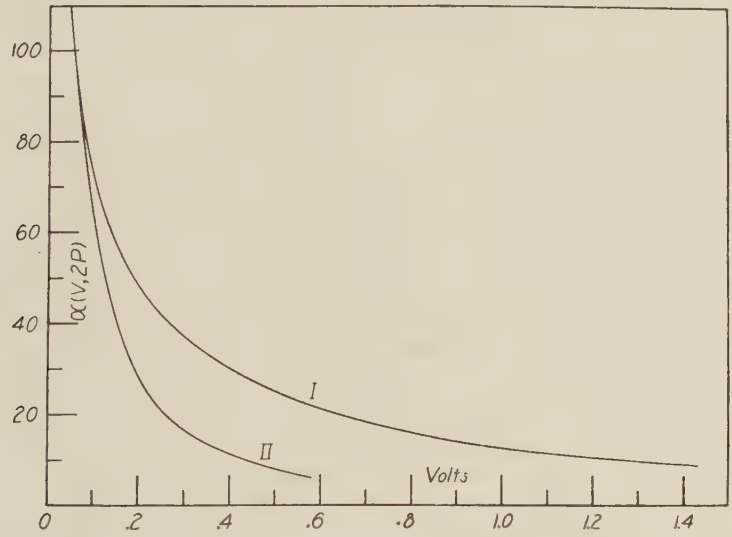


Fig. 13. Probabilities of recombination of an electron into the $2P$ states of caesium as a function of electron kinetic energy in volts. (Calculated from intensities of recombination spectra I; from photo-ionization II.)

calculated from spectral data, and that such experimental discrepancies as exist may be due to insidious errors which are very difficult to eliminate in work requiring such refined technique. The experimental work up to the present has, however, only just begun to give the information which may ultimately be secured. While others[206] have made interesting qualitative observations, Mohler alone has obtained all the information necessary to calculate recombination probabilities by this method.

[203] Milne, Phil. Mag. **47**, 209 (1924).

[204] Compton, "X-rays and Electrons," pp. 189–198. Oppenheimer, Zeits. f. Physik **41**, 268 (1927). Note that the absorption coefficient is proportional to ν^{-3}. The energy absorbed in one absorption act is $h\nu$. Thus the probability of absorption varies as ν^{-4}.

[205] Mohler, Foote and Chenault, Phys. Rev. **27**, 37 (1926).

[206] Paschen, Berlin Berichte, p. 135 (1926); Herzberg, Ann. d. Physik **84**, 553, 565 (1927); Mierdel, *ibid.* **85**, 612 (1928); Balasse, Comptes Rendus **184**, 1002, 1927 (1927).

In Fig. 13 are given experimental values of the relative probabilities of recombination into the $2P$ state in Cs as a function of the average electron energy. The absolute scale is given roughly by the estimate that $\alpha = 5.4(10)^{-14}$ at $V = 0.2$ volt. This value means that about 1 collision in 5000 results in recombination into the $2P$ state, for 0.2 volt electrons. Curve I, determined from the recombination spectrum, is in very close agreement with Eq. (67). Curve II is calculated from measurements on photo-ionization. The cause of the discrepancy between these curves is not understood.

Further theoretical work on the problem of recombination may only be mentioned. Milne[207] was led by a statistical equilibrium argument to conclude that, for small electron velocities, the effective combining cross section $q(v)$ must be of the form C/v^2. Since this is related to the absorption coefficient by $\alpha(v) = v \cdot q(v)$ we have

$$\alpha(v) = C/v. \tag{68}$$

If v is expressed in terms of ν and ν_i by Eq. (65) we have directly Eq. (67) except for the omission of the factor $\nu(\text{or } \nu^2)$, which is practically a constant for small electron velocities, where ν only slightly exceeds ν_i.

This also is in satisfactory agreement with the conclusions of Morse and Stueckelberg,[208] who have calculated $\alpha(v\ \nu_i)$ by wave mechanics and found, for every state, $\alpha(v\ \nu_i) = C/v = D/(\nu - \nu_i)^{1/2}$ as the asymptotic value for low velocities. Further, they obtained actual numerical values of $C(\text{ or } D)$ for the lower states and showed that the sum over all states is finite and varies as $1/v$ but cannot be expressed in simple form.

Kramers,[209] and Eddington[210] and Rüchardt[211] have derived theories for the combination of free electrons with stripped atomic nuclei, and that of Kramers contains only fundamental constants. The two former theories find their most interesting applications in X-rays and theories of the interior of stars; the latter has been applied to describe the neutralization of canal rays.

It may be noted in passing that three devices have proven particularly advantageous in enhancing the spectrum due to recombination. These are: (1) Schüler cathode,[212] a hollow cathode in a glow discharge from within which is emitted light characteristic of recombination (Paschen[206]); (2) the electrodeless ring discharge, from whose center the recombination light is emitted (Herzberg, Mierdel[206]); (3) a low voltage arc between a hot cathode of great electron emissivity and very small dimensions and a surrounding anode of large area. All of these arrangements are particularly favorable

[207] Milne, Phil. Mag. **47**, 209 (1925); see also Becker, Zeits. f. Physik **18**, 325 (1923).

[208] Morse and Stueckelberg, Phys. Rev.; **35**, 116 (1930); see also Oppenheimer, Zeits. f. Physik **55**, 725 (1929), who developed a method and obtained general equations by wave mechanics, but whose results were not worked out for the particular problem here considered.

[209] Kramers, Phil. Mag. **46**, 836 (1923; Wentzel, Zeits. f. Physik **27**, 257 (1924).

[210] Eddington, Inner Constitution of the Stars.

[211] Rüchardt, Zeits. f. Physik **15**, 164 (1923).

[212] Schüler, Zeits. f. Physik **35**, 323 (1926); **37**, 728 (1926).

for the production of large electron and ion concentrations in regions of such small electric field that the electron velocities are low.

Finally, the *capture of electrons by swiftly moving α particles* has very recently been investigated by Davis and Barnes[213] with most interesting and surprising results. It has been known[214] that fast α particles do not capture electrons from the molecules through which they pass until their velocity has decreased below a certain limiting value, suggesting that capture is impossible if the relative velocity exceeds this critical value.[215] For example it was found that α particles whose original velocity was 2.06 $(10)^9$ cm/sec. did not capture any electrons until their speed had fallen to about 0.82 $(10)^9$ cm sec. and did not capture the second electron to become completely neutralized unless the velocity were less than about 0.31 $(10)^9$ cm/sec. In order to examine this point more closely, Davis and Barnes projected a beam of α particles into an evacuated vessel in which also a stream of electrons from a hot cathode was passed with regulated speed in the same

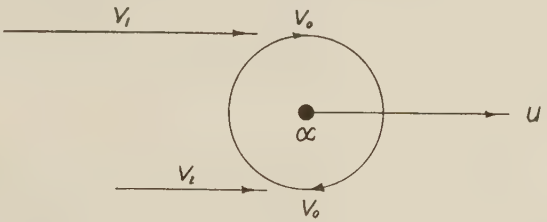


Fig. 14. Conditions for capture of electrons by α particles. u=velocity of α particle; v_0=orbital velocity which an electron would have in any one of the Bohr orbits of He^+; v_1 and $=v_2$ velocities of free electrons.

direction. Thus each α particle found itself accompanied in its onward flight by a cloud of electrons whose speed, relative to itself, could be adjusted at will. Particles which had captured electrons were then sorted out from the rest by action of a magnetic field, and the number of particles which did not capture electrons was counted on a fluorescent screen.

With this arrangement it was found that there was a relatively high probability of capture when the relative velocity of electrons and α particles had certain definite characteristic values, and that these characteristic velocities were related to the orbital velocities of electrons in He^+ in the simple way illustrated in Fig. 14. The condition for capture of a free electron by an α particle is that $v_0=v_1-u$ or $v_0=u-v_2$. In other words, the relative velocity between electron and α particle must equal the velocity which an electron would have in an orbit, if captured. But this relative velocity refers to conditions *before* the electron is accelerated toward the α particle by the Coulomb field in the process of its capture. Thus another way of express-

[213] Davis and Barnes, Phys. Rev. **34,** 152 (1929); Barnes, *ibid.* **34,** 1224 (1929); **35,** 217 (1930).

[214] Henderson, Proc. Roy. Soc. **A102,** 496 (1923).

[215] Davis, Nature, May 26, 1923.

ing the situation is to say that capture is probable if the total energy relative to the α particle is twice the energy in some one of the Bohr orbits. In addition to this, capture is possible if the relative velocity is zero. Similar energy relations hold for the simultaneous capture of two electrons, with the interesting fact that the condition for recombination appears to be simply that the sum of the energies of the two electrons must equal the total negative energy of the final state of the helium atom.

There is as yet no satisfactory explanation of these experiments. That the experiment should have yielded any results at all is surprising in view of the fact that, to effect the captures which were found, it is necessary that an electron with suitable velocity is captured if its trajectory carries it within a distance of the order of 10^{-5} cm from the α particle,—a distance much larger than the kinetic theory atomic radius. In other words, these experiments indicate an effective atomic cross section for recombination of the order of a million times larger than the ordinary kinetic theory atomic cross section, for the specified relative energies including zero relative energies. For the recombination when the relative velocity is close to zero this is quite to be expected according to the above theories, but recombination at other energies appears to involve quite a different process from any at present recognized.[216]

D. Motions of Electrons and Ions

(1) **Classical kinetic theory** will be understood as referring to that type of treatment of particle kinetics which has been used with considerable success in the problem of molecular motions in the kinetic theory of gases. We shall first outline the more important relations which can thus be derived, then consider experimental tests of this classical theory, and finally mention some attempts to apply the new methods of quantum mechanics to the problem.

(a) *Maxwell's distribution of velocities*,[217] so well known in kinetic theory, is found experimentally to describe very closely the motions of electrons in most regions of a discharge, and we shall see also that there is every theoretical reason for expecting it also to describe motions of ions except in extreme circumstances. The distribution law may be variously expressed in terms of the number of particles per unit volume N:

$$Nf(u)du = N\left(\frac{hm}{\pi}\right)^{1/2} \epsilon^{-hmu^2} du \tag{69}$$

is the number with velocity components between u and $u+du$ in a given direction;

$$Nf(uvw)dudvdw = N\left(\frac{hm}{\pi}\right)^{3/2} \epsilon^{-hm(u^2+v^2+w^2)} dudvdw \tag{70}$$

[216] Stueckelberg and Morse, Phys. Rev. **35**, 116 (1930); Adams, *ibid.* **34**, 537 (1929).
[217] Jeans, "Dynamical Theory of Gases," 2 ed. Chap. II.

is the number with perpendicular velocity components lying between u and $u+du$, v and $v+dv$, w and $w+dw$;

$$NF(cdc)=4\pi N\left(\frac{hm}{\pi}\right)^{3/2}\epsilon^{-hmc^2}c^2dc \tag{71}$$

is the number with speeds (irrespective of direction) lying between c and $c+dc$. By integration of this equation between any assigned value c and infinity, one can find the number of particles with velocities greater than c. Furthermore,

$$Nuf(u)du=N\left(\frac{hm}{\pi}\right)^{1/2}u\epsilon^{-hmu^2}du \tag{72}$$

is the number *which cross any plane* with velocity components perpendicular to the plane lying between u and $u+du$.

The constant h may be written in terms of the average kinetic energy $\overline{E}$, or the temperature T, or the mean square molecular speed C^2 by

$$\overline{E}=\frac{1}{2}mC^2=\frac{3}{2}kT=\frac{3}{4h} \tag{73}$$

where k is the Boltzmann gas constant.

By integration of Eq. (72) over all values of u, the total number of particles crossing unit area in either direction is, per second

$$n=N\frac{1}{2(\pi hm)^{1/2}}=N\left(\frac{kT}{2\pi m}\right)^{1/2}. \tag{74}$$

The average speed $\bar{c}$ and the root mean square speed C are related by

$$\bar{c}=(8/3\pi)^{1/2}C=0.921C. \tag{75}$$

Comparing Eqs. (72 and 74) it is seen that

$$u\epsilon^{-hmu^2}du \tag{76}$$

represents the fraction of the particles crossing a surface which possess velocity components perpendicular to the surface lying between u and $u+du$. From this, we find that the average energy of those particles *which cross a surface* is $2\ kT$, whereas the average energy of the particles *in a given volume* is $(3/2)kT$. The difference is due to the fact that, for the particles which cross a surface, the average energy associated with each of the two coordinates parallel to the surface is the usual value $1/2\ kT$, but that associated with the coordinate perpendicular to the surface is kT. This increased value is due to the fact that relatively more fast particles than slow ones cross the surface.

If any two regions are separated by a layer in which there is a force field of such a nature that a definite amount of work W would be done by every particle in passing from the first region to the second, then it can be

shown by direct integration[218] of Maxwell's equations that, if the particles in the first region possess a Maxwell's velocity distribution, those which pass through the retarding layer enter the second region also with a Maxwellian distribution characteristic of the same temperature. Indeed, by *Boltzmann's equation*,[219] the only effect of such a force field is to alter the concentrations of the particles in the two regions in the ratio

$$\frac{n_1}{n_2} = \epsilon^{-W/kT} \tag{77}$$

where W is the difference in the potential energies of a particle in the two regions.

Mott-Smith and Langmuir[220] have proven an analogous theorem which applies to a more general case of particles with a Maxwellian distribution of velocity which cross a surface into a region of *either retarding or accelerating* field toward some second surface, and which is applicable to problems of collecting electrodes. It shows that in this case also the Maxwell-Boltzmann equations apply to the particles reaching the second surface, *but only to certain groups in such a distribution*. Consider a conservative system consisting of a large number of particles moving in an enclosed space and continually exchanging energy and momentum with each other so that the state of statistical equilibrium described by the Maxwell-Boltzmann distribution law is reached. In this space we imagine to exist a region A in which there is no interaction by collision or otherwise between the particles, and in which there is a field of force acting on the particles. Any interior boundary surfaces of A (such as surfaces of collectors) we shall first assume to be perfectly reflecting. The particles penetrating into A from the exterior will then describe "orbits" (perhaps including reflections at the interior surfaces) and eventually escape again across the exterior boundary. There may, in addition, be some orbits within A which never intersect the outer boundary. This would be true, for example, if there were a force directing the particles inward into A, and we were to introduce particles within A with kinetic energies too small to permit their escape. Now consider all possible orbits, such as we would have if we imagined particles to pass every point within A with every velocity between zero and infinity. Out of this group of orbits, some will intersect the outer boundary. Others *may* be present (as in case of a field accelerating particles inward) which lie entirely within the space A. These paths which never carry the particles outside of A we shall call "interior orbits." The theorem is now as follows:

"*If the field A has no interior orbits, the distribution of particles and their velocities throughout it will conform to the Maxwell-Boltzmann distribution law. If interior orbits exist, the distribution will also conform to the Maxwell-Boltzmann law except for an excluded class of particles, these being the particles which, if present, would describe the interior orbits.*"

[218] Richardson, Phil. Mag. **18,** 695 (1909).

[219] See, for example, Richardson, "Electron Theory of Matter," Chap.

[220] Mott-Smith and Langmuir, Phys. Rev. **28,** 727 (1926); see especially pp. 756–758.

The first case includes the case of electrons or ions which cross the boundary of a plane or convex sheath within which they are retarded and drawn back to the boundary. The second case is illustrated by electrons or ions entering a sheath within which they are accelerated toward an interior collecting electrode or repelled back to an enclosing surface, as in a hollow collector.

If the interior surfaces are absorbing instead of perfectly reflecting, then the distribution is modified in that the group of particles which would have velocities coming from that surface is missing. The simplest case is that of a plane collector surrounded by a space charge sheath with a parallel boundary. Here the concentration of particles of such speeds as would carry them to the collector is just half what it would be if the collector were perfectly reflecting. For the particles which are *drawn* in, there is just a half (or "one way") Maxwell distribution after exclusion of all particles which would have interior orbits. For the particles which penetrate *against* the field within the sheath, there is everywhere the full Boltzmann distribution for those particles which cannot penetrate to the inner absorbing surface, and a half Boltzmann distribution for those which can so penetrate. There are numerous possible applications of this theorem to the condition of ions in the neighborhood of electrodes of various shapes and at various potentials with respect to the surrounding ionized gas.

(b) *The mean free path*[221] is the average distance moved by a particle between collisions. For real molecules its value obviously depends on our definition of a *collision*. For hard spherical particles, which are so often discussed in kinetic theory as a convenient approximation, there is no question; two particles (1 and 2) collide whenever their centers approach each other to within the distance $\sigma_{12}=\sigma_1+\sigma_2$, the sum of their radii. The mean free path λ_1 of particles of type 1 which collide with particles of type 2 is then

$$\lambda_1=\frac{1}{\pi N_2\sigma_{12}^2(1+h_1m_1/h_2m_2)^{1/2}}=\frac{1}{\pi N_2\sigma_{12}^2(1+C_2^2/C_1^2)^{1/2}}$$
$$=\frac{1}{\pi N_2\sigma_{12}^2(1+\bar{E}_2m_1/\bar{E}_1m_2)^{1/2}} \tag{78}$$

provided both types of particles possess Maxwellian velocity distributions characteristic of mean energies $\bar{E}_1$ and $\bar{E}_2$ respectively,—or velocities C_1 and C_2. In problems of kinetic theory of gases, the mean energies of admixed gases are equal, and hence the factors h_1/h_2 and E_2/E_1 drop out of Eq. (78), but this is not true of electrons or ions in gas discharges. Eq. (78) reduces to simpler forms in important special cases:

$$\lambda_1=\frac{1}{2^{1/2}\pi N_2\sigma_{12}^2} \tag{79}$$

[221] Jeans, "Dynamical Theory of Gases," 2 ed., Chap. X.

if both particles are of the same type;

$$\lambda_1 = \frac{1}{\pi N_2 \sigma_{12}^2} \tag{80}$$

irrespective of velocity distributions if the velocity of type 1 is much greater than that of type 2, as in the case of electrons colliding with gas molecules;

$$\lambda_1 = \frac{1}{4/(2)^{1/2}\pi N_2 \sigma_2^2} \tag{81}$$

if the collision radius σ_2 of the target particle so exceeds that of the projectile particle that it may be substituted for the sum of the radii of the colliding particles in Eq. (79).

If σ_2 is the radius of a molecule, a comparison of Eqs. (80, 81) shows that

$$\lambda_{\text{electron}} = 4(2)^{1/2}\lambda_{\text{molecule}}, \tag{82}$$

since the contribution of the electron radius σ_1 to the sum σ_{12} for electron and molecule is negligible. Eq. (82) defines the so-called "*kinetic theory mean free path of an electron.*" Some values thus calculated from accepted values of molecular free paths are shown in Table XV, together with their reciprocals $\nu = 1/\lambda$ which give the average numbers of collisions per centimeter path.

TABLE XV. *Kinetic theory mean free paths* λ *and numbers of collisions* ν *per centimeter path in gas at pressure* $p = 1$ *mm at 25°C.*

Gas	λ (electron)	λ (molecule)	ν (electron)	ν (molecule)
Hg	0.0149 cm	0.00263 cm	67.0	380.0
A	0.0450	0.00795	22.2	125.9
Ne	0.0787	0.01390	12.7	72.0
He	0.1259	0.02221	7.95	45.0
H_2	0.0817	0.01444	12.2	69.1
N_2	0.0425	0.00751	23.5	133.0
O_2	0.0455	0.00805	22.0	124.2
HCl	0.0322	0.00570	31.0	175.6
CO	0.0420	0.00743	23.8	136.5

Closely related to these expressions is[187]

$$\nu_{12} = 2\pi^{1/2} N_1 N_2 \sigma_{12}^2 \left(\frac{1}{h_1 m_1} + \frac{1}{h_2 m_2}\right)^{1/2} \tag{83}$$

for the total number of collisions per unit volume per unit time between particles of type 1 and those of type 2.

When the particle whose free path we are considering is an ion, there is an attraction between it and all neighboring uncharged particles due to the electric doublet induced in each. The magnitude of this force between an ion and a neutral molecule distant r is,

$$F=\frac{K-1}{4\pi N}\frac{e^2}{r^5}=Ar^{-5} \tag{84}$$

where K is the dielectric constant and N is the number of molecules per unit volume.[222] This attraction both increases the number of times that the center of an ion comes within the distance σ_{12} of the center of a molecule, and also deflects the path of the ion when ion and molecule pass near each other, but not near enough to collide. Each of these effects reduces the diffusion and mobility of the ion. Since the diffusion constant D and mobility constant μ are proportional to the mean free path in the absence of attraction, we may use their reduction to define the mean free path in the presence of attraction, *i.e.*,

$$\frac{\lambda'}{\lambda}=\frac{\mu'}{\mu}=\frac{D'}{D}$$

where the primed symbols represent the values in the presence of attraction and the unprimed symbols represent the values which would be obtained if there were no attraction. Langevin[223] has calculated μ and D for smooth elastic sphere ions moving among smooth elastic sphere molecules both with and without attraction when the ions and molecules are in thermal equilibrium, and hence have the same average kinetic energies $\bar{E}$. The effect of attraction is found to depend on the ratio $A/4\sigma_{12}{}^4\bar{E}$. $A/4\sigma_{12}{}^4$ is the work required to separate an ion and molecule from contact to infinity against the force of Eq. (83). It is therefore the *energy of dissociation of an elementary cluster ion.* The following table shows the magnitude of the effect:

TABLE XVI.

$\frac{A/4\sigma_{12}{}^4}{\bar{E}}$	$\frac{\lambda'}{\lambda}=\frac{\mu'}{\mu}=\frac{D'}{D}$
∞	0
10	0.076
5	.15
2.5	.311
1.0	.732
0.5	.93
0.3	.97
0.0	1.00

Had we defined mean free paths by energy loss instead of by diffusion and mobility, these values would be somewhat different, but the general conclusion would remain unchanged. This conclusion is that *the ion free path is not appreciably diminished by the fact of its charge unless the mutual kinetic energy is of the order of the cluster dissociation energy, or less.* Hence in ionized gases in which ion clusters do not occur to an appreciable extent, we do not

[222] Langevin, Ann. d. Chem. et de Phys. **28**, 316 (1903).
[223] Langevin. Ann. d. Chim. et de Phys. **5**, 245 (1905); Hassé, Phil. Mag. **1**, 139 (1926).

need to consider the effect of the ion charge upon its free path. Apparently then this phenomenon plays no appreciable role in ordinary vacuum discharge tubes or any discharges which dissipate large amounts of energy density. It may be important in phenomena of feeble electrical conduction through gas at considerable pressure.

(c) *Collisions and single scattering* are most simply, and for many purposes rather satisfactorily, handled in the kinetic theory of gases by considering collisions to be like those of elastic spheres.

Energy losses at collisions are easily calculated for impacts of *elastic sphere.* Assume a target sphere of mass M at rest before being struck at random by a projectile sphere of mass m and original velocity v. Let θ be

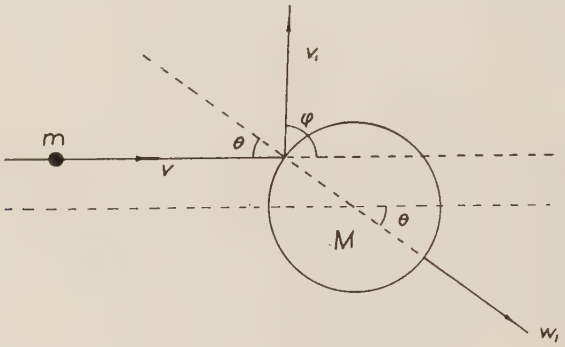


Fig. 15. Collision of elastic spheres.

the angle between the original trajectory of m and the radius vector in M to the point of impact. The results of the impact are determined by the momentum and energy equations

$$\begin{aligned} mv - mv_1 \cos\phi &= Mw_1 \cos\theta \\ mv_1 \sin\phi &= Mw_1 \sin\theta \\ \tfrac{1}{2}mv^2 - \tfrac{1}{2}mv_1^2 &= \tfrac{1}{2}Mw_1^2 \end{aligned} \tag{85}$$

from which

$$f_\theta = \frac{\tfrac{1}{2}mv^2 - \tfrac{1}{2}mv_1^2}{\tfrac{1}{2}mv^2} = \frac{4Mm}{(M+m)^2}\cos^2\Theta \tag{86}$$

is the fraction of its initial energy which is lost by the projectile particle at this collision specified by θ. The projectile particle may strike the target particle at any point of its surface specified by θ between 0 and $\pi/2$. Multiplying f_θ by the probability of striking at an angle between θ and $\theta+d\theta$, and integrating over all values of θ from 0 to $\pi/2$ gives the *average fraction of energy lost at a collision*

$$f = \frac{2Mm}{(M+m)^2}\,. \tag{87}$$

From this theory we predict that electrons impinging against molecules lose, on the average, a fraction $f=2m/M$ of their energy (where $m \ll M$). Ions, on the other hand, colliding with molecules of the same mass, are expected to lose one-half their energy, on the average.

When the target particles are also in motion, the average fraction of energy lost by the impinging projectile particles is less than that given by Eq. (87). If the projectile particle is moving quite rapidly and has small mass m compared with the target particle M, Compton[224] showed that the average fraction of energy lost by a projectile particle at impact is

$$f=2\frac{m}{M}\left(1-\frac{E_m}{E_M}\right) \tag{88}$$

where E_m and E_M are the kinetic energies of the projectile and target particles, respectively. This expression should be applicable to electrons colliding with molecules (as elastic spheres) provided the velocities of each of the two types of particle were homogeneous.

More recently Cravath[225] has succeeded in solving the analogous problem for the general case of particles with Maxwellian distributions and with no restrictions upon the masses or average energies of the two groups. He finds

$$f=2.66\frac{mM}{(m+M)^2}\left(1-\frac{E_m}{E_M}\right) \tag{89}$$

for the *average fraction of the average energy* E_m lost by particles m at a collision with one of the particles M whose average energy is E_M. Thus the previous applications of Eq. (88) to problems of electron mobility should be corrected by substituting a factor 2.66 in place of 2.

Single scattering is easily handled in case the colliding particles behave like *elastic spheres*. From Eqs. (85) we find a general relation between the angle of incidence θ and the angle of scattering ϕ,

$$m[\sin(\phi-\theta)\sin(\phi+\theta)+\sin^2\theta]=M[\sin^2(\phi-\theta)-\sin^2\theta]. \tag{90}$$

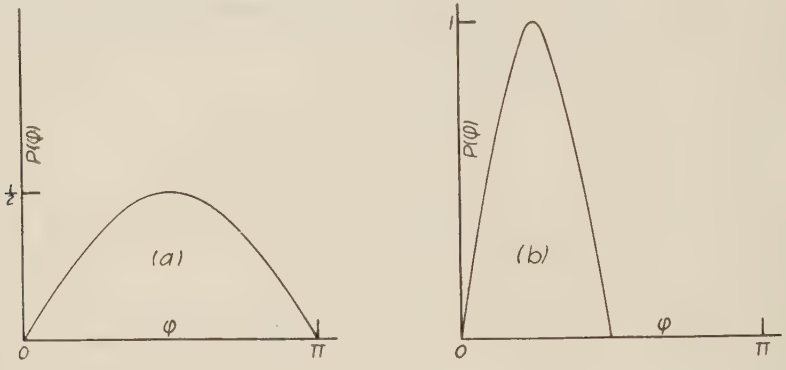
The probability of colliding at an angle between θ and $\theta+d\theta$ is $2\sin\theta\cos\theta\, d\theta$, whence the probability of deflection through an angle between ϕ and $\phi+d\phi$ is found by substitution of ϕ for θ in this expression according to their relation in Eq. (90). For the two most important cases of impacts of electrons against molecules ($m \ll M$) and impacts of ions against molecules of equal mass ($m=M$) we thus find for the relation between θ and ϕ, and for the probability $P(\phi)\,d\phi$ that a particle which collides is deflected through an angle between ϕ and $\phi+d\phi$

$$\begin{aligned}&\text{for electron } \phi=2\theta \text{ and } P(\phi)d\phi=\tfrac{1}{2}\sin\phi\, d\phi\\ &\text{for ion } \phi=\pi/2-\theta \text{ and } P(\phi)d\phi=\sin 2\phi\, d\phi.\end{aligned} \tag{91}$$

[224] Compton, Phys. Rev. **22,** 333 (1923).

[225] Cravath, to be published, probably in the Physical Review 1930.

We see, among other things, that the most probable deflection of an electron is at 90° and of an ion is at 45°; also that no ions are deflected backwards, and that the paths of the ion and molecule after collision are at right angle to each other. The distributions of scattering are shown graphically in Fig. 16. By referring to Eq. (86) one can easily find the energy transmitted to the molecule for any angle of impact or of deflection.



(a) electron impact (b) ion impact

Fig. 16. Distribution of scattering angles for elastic spheres.

In these expressions no distinction is made in regard to the azimuth of deflection, all azimuths at a given angle ϕ contributing to $P(\phi)$. It is sometimes desirable to express these distributions differently, in terms of the relative probabilities of deflection within a given element of solid angle for various orientations of this element about the axis of the original trajectory. In other words, if radius vectors were drawn from a common point for every collision, what would be the density of these vectors in different directions. This is readily found from Eqs. (91) by dividing each value of $P(\phi)\,d\phi$ by the element of solid angle included between ϕ and $\phi+d\phi$, *i.e.*, by $2\pi \sin\phi d\phi$. Calling the result $F(\phi)$, meaning the fraction of deflections per unit solid angle in the direction ϕ, we have:

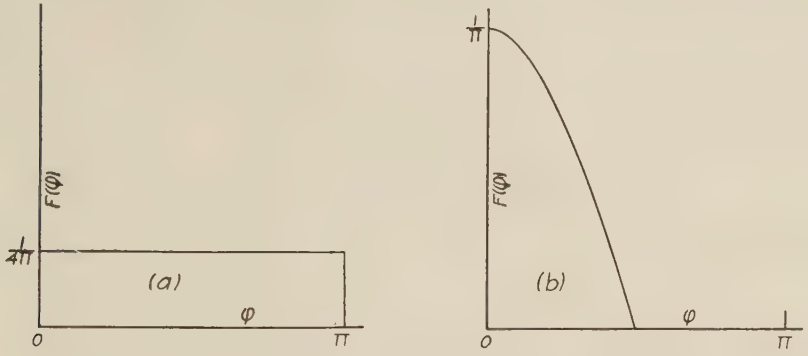
$$\text{for electrons } F(\phi)=1/4\pi \text{ ; for ions } F(\phi)=(\cos\phi)/4 \qquad (92)$$

showing that electrons are scattered equally in all directions in space, while ions are scattered most densely in the forward direction. Fig. 17 shows these distributions.

A more general classical case is that in which the particles exert on each other forces of the type $F \propto r^{-n}$. In this case there is no natural definition of a collision. The relative scattering at different angles can still be calculated, but this is now a function of the number and concentration of the scattering particles and of the velocities of the scattered particles. As an example consider:

Scattering of electrons by centers of inverse square force fields. Assume that an electron of kinetic energy $(1/2)mC^2=eV$ traverses a region in which

there are N singly charged molecular ions per unit volume. The attraction of the electron to each ion is e^2/r^2. In a path of length l through this ionized



(a) electron impact (b) ion impact

Fig. 17. Alternative graph of distribution of scattering angles.

gas the probability of deflection through an angle between ϕ and $\phi+d\phi$ as the result of the interaction with some one positive ion is[226]

$$P(\phi)d\phi=\frac{\pi e^2 Nl}{4V^2}\cot\frac{\phi}{2}\cdot\csc^2\frac{\phi}{2}d\phi \tag{93}$$

and the fraction of deflections per unit solid angle in the direction ϕ is

$$F(\phi)=\frac{e^2 Nl}{16V^2}\csc^4\frac{\phi}{2}\cdot \tag{94}$$

If the energy is expressed as V volts, the coefficients of the trigonometric quantities reduce to 16.1 $(10)^{-5}$ Nl/V^2 and 1.28 $(10)^{-15}$ Nl/V^2, respectively These two distribution functions are illustrated in Fig. 18.

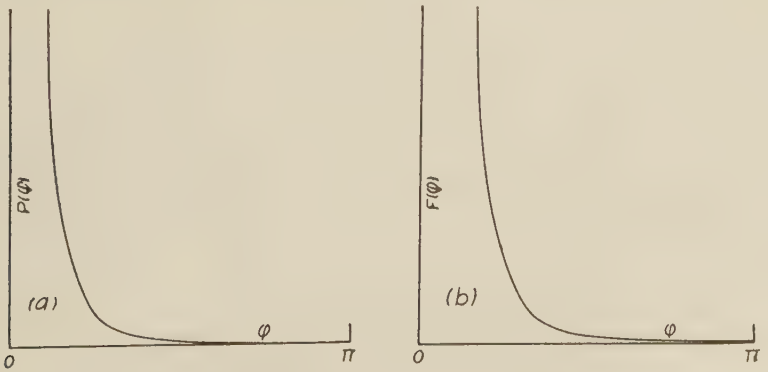


Fig. 18. Single scattering of electrons by ions (inverse square law).

[226] This is the well-known formula for single scattering as applied by Darwin to the scattering of α particles (Rutherford, Phil. Mag. **21**, 669 (1911); Millikan, "The Electron," Appendix F). Although in these references the scattering is worked out for an inverse square law of *repulsion*, the same method applied to the problem of *attraction* leads to the identical result.

Since collisions of elastic spheres are equivalent to deflections by centers of a force $F \propto r^{-\infty}$, it is obvious that for any other inverse power law of force higher than the second, the distribution functions would be intermediate in character beween those illustrated in Fig. 18 and those in Figs. 16a and 17a.

Finally, if we investigate the single scattering of *ions* by N ions per unit volume of the same mass we find,[227] in path

$$P(\phi)d\phi = \frac{2\pi e^2 Nl}{V^2} \cot\phi \csc^2\phi \, d\phi \tag{95}$$

$$F(\phi) = \frac{e^2 Nl}{V^2} \cot\phi \csc^3\phi$$

shown graphically in Fig. 19. With V in volts, the coefficients here are 129 $(10)^{-15} Nl/V^2$ and 20.5 $(10)^{-15} Nl/V^2$, respectively.

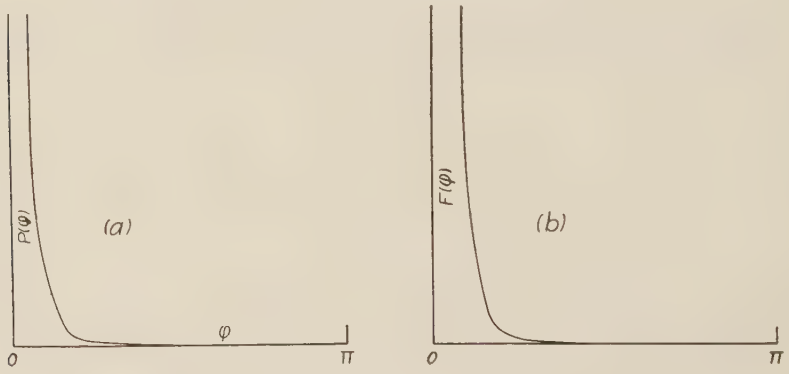


Fig. 19. Single scattering of ions by ions (inverse square law).

It will be observed that the scattering by elastic spheres differs in one important detail from that by force centers in that the expressions for $P(\phi)$ and $F(\phi)$ do not involve the concentration N of scattering centers. This is because the definition of a collision in the former case is perfectly precise, whereas in the latter case there is no distinction between a collision and a non-collision. Really to compare different types of scattering it is preferable never to speak of scattering at a collision, but rather of single scattering per unit path through the gas containing N molecules per unit volume. Then expressions (93, 94, 95) remain unchanged, but for elastic spheres we must modify Eqs. (91, 92) thus: The aggregate cross sectional area for collision offered by the N molecules in unit volume is $\pi\sigma_{12}^2 N$. In path l the chance of collision is $\pi\sigma_{12}^2 Nl$ and of non-collision is $(1 - \pi\sigma_{12}^2 Nl)$.

[227] By an adaptation of the methods developed for study of the recoil of light atoms at collisions with α particles. (Darwin, Phil. Mag. **27**, 499 (1914); Rutherford, *ibid.* **37**, 537 (1919); Proc. Roy. Soc. **97**, 374 (1920).

The former fraction of particles traversing path l is scattered according to Eqs. (91, 92). The latter fraction is not scattered at all. Thus we have: for electrons:

$$P(\phi)d\phi = \begin{cases} (1-\pi\sigma_{12}^2 Nl) \ldots\ldots\ldots\ldots\ldots \phi = 0 \\ \pi\sigma_{12}^2 Nl \cdot \tfrac{1}{2} \sin\phi \, d\phi \ldots\ldots\ldots 0 < \phi < \pi \end{cases}$$

$$F(\phi) = \begin{cases} \infty \ldots\ldots\ldots\ldots\ldots\ldots\ldots \phi = 0 \\ \dfrac{1}{4\pi}\pi\sigma_{12}^2 Nl \ldots\ldots\ldots\ldots\ldots 0 < \phi < \pi \end{cases} \tag{96}$$

for ions:

$$P(\phi)d\phi = \begin{cases} (1-\pi\sigma_{12}^2 Nl) \ldots\ldots\ldots\ldots\ldots \phi = 0 \\ \pi\sigma_{12}^2 Nl \cdot \sin 2\phi d\phi \ldots\ldots\ldots 0 < \phi < \pi/2 \end{cases}$$

$$F(\phi) = \begin{cases} \infty \ldots\ldots\ldots\ldots\ldots\ldots\ldots \phi = 0 \\ \pi\sigma_{12}^2 Nl \cdot \dfrac{1}{\pi} \cos\phi \ldots\ldots\ldots\ldots 0 < \phi < \pi/2. \end{cases} \tag{97}$$

Fig. 20 illustrates the application of these expressions to the case of single scattering of electrons in helium. The total area under each curve must, of course, equal unity. In the case of scattering by elastic spheres, the area $OABC$ equals the fraction which have collided, and the line CD goes to infinity in such a way that the area under it has a finite area such that $OABCD$ equals unity. For any inverse power law higher than the second, the curves lie intermediate between the two types shown.

(d) *Diffusion* is one of the two processes by which ions move in a discharge, the other process being the action of the field. The fundamental law of diffusion is[228]

$$D_{12} = \frac{1}{3} \frac{N_1\lambda_2 c_2 + N_2\lambda_1 c_1}{N_1 + N_2} \tag{98}$$

for the interdiffusion of gases 1 and 2, whose molecular concentrations, mean free paths and average speeds are denoted by N, λ and c, respectively. The number of molecules of type 1 which cross unit area in the direction of z in unit time is

$$n_1 = -D_{12}\frac{\partial N_1}{\partial z} \tag{99}$$

and the rate of increase of their concentration in unit volume is

$$\frac{dN_1}{dt} = D_{12}\nabla^2 N_1. \tag{100}$$

[228] Jeans, "Dynamical Theory of Gases," 2 ed. p. 326.

Substituting the values of λ from Eq. (78) and putting $N=N_1+N_2$ for the total number of molecules per unit volume, we obtain

$$D_{12}=\frac{4}{(3\pi)^{3/2}N\sigma_{12}^2}\left(\frac{\overline{E}_1}{m_1}+\frac{\overline{E}_2}{m_2}\right)^{1/2}$$
$$=\frac{0.921(C_1^2+C_2^2)^{1/2}}{3\pi N\sigma_{12}^2}$$

where the average velocity c is converted in to r.m.s. velocity C by Eq. (75) and $\overline{E}$ is the mean kinetic energy by Eq. (7 3). (We may note that, in sub-

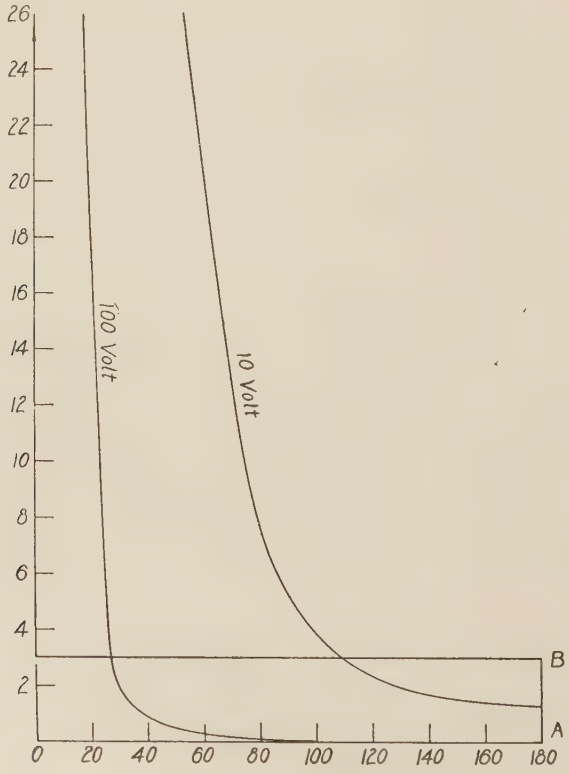


Fig. 20. Comparison of scattering as elastic spheres and as centers of inverse square force (electrons in helium). Ordinate is $F(\phi)/Nl(10)^{-17}$

stituting in Eq. (98) the values of λ_1, λ_2 given by Eq. (78), we have assumed that λ_1, λ_2 refer only to free paths of the *one* gas which are terminated by collisions with the *other* gas. We have neglected collisions between molecules of the same gas. This is legitimate, however, since collisions of this latter type do not, on the average, affect the rate of diffusion.)

In these equations let us introduce a quantity λ_0 defined by

$$\lambda_0=(\pi N\sigma_{12}^2)^{-1}. \qquad (102)$$

This is the free path which a particle would have if its speed were very large compared with the speeds of the target particles (Eq. (80)). Let us also limit ourselves to the special cases of electrons diffusing through the gas. ($m_1 \ll m_2$ and $C_1 \gg C_2$) and ions diffusing through the gas ($m_1 = m_2 = M$). Eq. (101) then reduces to

$$D_{\text{electrons}} = 0.434\lambda_0(\overline{E}_1/m_1)^{1/2} = 0.333\lambda_0 c_1$$
$$D_{\text{ions}} = 0.434\frac{\lambda_0}{M^{1/2}}(\overline{E}_1^2+\overline{E}_2^2)^{1/2} = 0.333\lambda_0(c_1^2+c_2^2)^{1/2}. \tag{103}$$

It is to be noted here that, for electrons, λ_0 is the "kinetic theory" mean free path; for ions the kinetic theory value of λ_0 is 1/4 of that for electrons owing to the negligible radii of the latter. λ_0 for ions is, however, $2^{1/2}$ times the molecular mean free path.

Equations (101, 103) are practically identical with those derived from direct consideration of "persistence of velocities," which phenomenon principally accounts for the departure of the numerical coefficient from the value 1/3 when the diffusing particles are of appreciable mass.[229] We have followed the Stefan-Maxwell method, which avoids the direct consideration of persistence of velocities as required by the alternative Meyer theory.

(e) *Mobility* μ is defined in terms of the electric field strength E and the average of ionic drift by the relation

$$\bar{v} = \mu E \tag{104}$$

which is useful even though μ is in general not a constant, but is a function of E. The value of μ has been deduced by a variety of methods,[230] whose results differ only in the numerical coefficient, whose value varies by a factor of about 2 between the extreme cases. A satisfactory method is to deduce it from the diffusion constant D as given in Eqs. (101) by use of the general relation

$$\mu = \frac{e}{kT}D \tag{105}$$

where $3kT/2$ is the average kinetic energy $\overline{E}$ of the ions. This relation was derived by Thomson[231] by equating the average rate of drift of ions across unit area caused by the electric field, $N\mu E$, to the drift which would be caused by a partial pressure gradient $-dp/dz$ sufficient to give the same force on the N ions of unit volume, *i.e.*, with $NeE = -dp/dz$. Hence $N\mu E = -(\mu/e)\,(dp/dz) = -(\mu kT/e)\,(dN/dz)$ in virtue of $p = NkT$. But this must be identical with the diffusion drift $-D\,dN/dz$, whence Eq. (105) follows immediately.

[229] Jeans, "Dynamical Theory of Gases," 2 ed., p. 328.

[230] Langevin, Ann. de Chim et de Phys. **5**, 245 (1905); **28**, 317, 495 (1903); Mayer, Ann. der Phys. **62**, 358 (1920); Loeb, "Kinetic Theory of Gases," Chap. XI; Townsend, "Electricity in Gases," p. 84.

[231] Thomson, "Conduction of Electricity Through Gases," 3 ed., p. 79.

In this manner Eqs. (73, 101, 105) lead to

$$\mu_{12}=\frac{2e}{(3\pi)^{1/2}\pi N\sigma_{12}^{2}\overline{E}_1}\left(\frac{\overline{E}_1}{m_1}+\frac{\overline{E}_2}{m_2}\right)^{1/2}$$
$$=\frac{0.921e}{\pi N\sigma_{12}^{2}\cdot m_1C_1}\left(1+\frac{\overline{E}_2m_1}{\overline{E}_1m_2}\right)^{1/2} \tag{106}$$

for the mobility of charged particles 1 moving through gas 2. If the average velocity c_1 is used instead of the r.m.s. velocity C_1, the numerical coefficient is 0.85 instead of 0.921.

Again introducing λ_0 from Eq. (102) we have in general

$$\mu_{12}=\frac{0.921e\lambda_0}{m_1C_1}\left(1+\frac{\overline{E}_2m_1}{\overline{E}_1m_2}\right)^{1/2}$$

for electrons

$$\mu=\frac{0.921e\lambda_0}{mC}=\frac{0.85e\lambda_0}{mc} \tag{107}$$

for ions

$m_1=m_2=M$ $$\mu=\frac{0.921e\lambda_0}{MC}\left(1+\frac{\overline{E}_2}{\overline{E}_1}\right)^{1/2}.$$

Comparing our derivation of μ_{12} with the most familiar mobility equation, that of Langevin[232]

$$\mu_{12}=\frac{0.815e\lambda_0}{m_1C_1}\left(1+\frac{m_1}{m_2}\right)^{1/2}, \tag{108}$$

we notice two differences. The difference in the numerical factor arises from differences in the methods of averaging, and is typical of the variety of constants lying between 0.5 and 1.0 which follow the use of different degrees of refinement. We believe that the value here given is as reliable as any. The absence of the ratio of the mean energies in Langevin's (and all other) mobility equations arises from an implicit assumption of equipartition of energy in carrying out an integration, and has probably crept in because, when mobility equations were first derived, interest in mobility centered on cases where the ions were practically in equilibrium with the gas molecules.

The thing in Eqs. (107, 108) which is generally misunderstood is that λ_0 is not necessarily the mean free path, but is defined by Eq. (102) and is $2^{1/2}$ times greater than the mean free path if the velocities of the ions and molecules are of the same order.[233]

[232] Langevin, Ann. de Chim. et de Phys. **5**, 245 (1905); **28**, 495 (1903); Mayer, Ann. d Phys. **62**, 358 (1920).

[233] Loeb (Kinetic Theory of Gases, p. 447) calls particular attention to the correct definition of λ_0.

It should be remarked that Langevin has derived a more general mobility equation based on the assumption of inverse power attractive forces between ions and molecules.[234] Since we have already seen that attractive forces do not notably affect free paths when the ion energies are considerably too great to permit formation of "clusters," this refinement does not appear to be important for the theory of electric discharges in gases.

To the extent to which collisions may be treated as between elastic spheres, Eq. (107) should permit an estimate of mobilities accurate within the small uncertainty in the numerical coefficient, *provided the mean velocities and energies are known*. In strongly ionized gases these quantities can be measured directly for electrons and estimated with considerable precision for ions as the result of the development of special probe-wire methods[235] which will be described in Part II. Where this direct information is unavailable, Eq. (107) can only be used in connection with some supplementary theory for estimating the velocities and energies. This has been attempted in the following analysis, based on papers by Hertz[236] and Compton.[237]

Terminal speeds may be calculated by equating the rate of gain to the rate of loss of energy as a charged particle moves through a gas in a uniform field E. In advancing a distance dx in the direction of the field, each particle of charge e receives energy $edU = eEdx$, where U is the kinetic energy of the charged particle in terms of equivalent potential drop. In the same path dx the particle will lose energy $\nu' \cdot feU\, dx$, where ν' is the average number of collisions made while advancing unit distance and f is the average fraction of its energy lost at a collision as a result of momentum transfer. Thus

$$e\, dU/dx = e(E - \nu' f U) \tag{109}$$

is the net rate at which the particle acquires additional energy. The state of terminal speed is defined by $dU/dx = 0$, hence by the terminal energy U_t in

$$U_t = E/\nu' f. \tag{110}$$

From this the average number of collisions ν' made by a charged particle while advancing unit distance may be calculated thus:

In Eq. (107) put $eU = \bar{E}_1$ and $e\Omega = \bar{E}_2$ for the mean energies of charged particle and gas molecule, respectively. Then we have the average velocity of advance in the direction of the field E given by

$$\bar{v} = \mu E = \frac{0.921 e\lambda_0}{mC} E \left(1 + \frac{\Omega m}{UM}\right)^{1/2}. \tag{111}$$

[234] Langevin, Ann. de Chim. et Phys. **8,** 238 (1905); Thomson, "Conduction of Electricity Through Gases," 3 ed., p. 165.

[235] Langmuir and Mott-Smith, G. E. Rev. **27,** 449, 538, 616, 762, 810 (1924); also Tonks and Langmuir, Phys. Rev. **34,** 877 (1929).

[236] Hertz, Zeits. f. Physik **32,** 298 (1925); Physica, **2,** 15 (1922).

[237] Compton, Phys. Rev. **22,** 333 (1923).

But the average number of collisions made by a charged particle in unit time is c/λ. Hence the average number of collisions made while advancing unit distance is

$$\nu' = \frac{c/\lambda}{\mu E} = \frac{0.921C}{\lambda E} \frac{mC}{0.921e\lambda_0\left(1+\frac{\Omega m}{UM}\right)^{1/2}}$$
$$\nu' = \frac{2U}{\lambda\lambda_0 E\left(1+\frac{\Omega m}{UM}\right)^{1/2}} = \frac{2U}{\lambda_0^2 E} \tag{112}$$

by Eq. (78). Remembering that $1/\lambda_0$ is of the order of the number of collisions per unit path, we see that the number of collisions made while *advancing* unit distance is of the order of the square of the number made in unit path *along the trajectory*.

Returning now to Eq. (110) and substituting for f and ν' from Eqs. (89, 112) we obtain for the terminal energy of charged particles of mass m and mean free path λ in a medium whose molecules have mass M and average energy Ω,

$$U_t = \frac{\Omega}{2} + \left[\frac{\Omega^2}{4} + \frac{\lambda_0^2 E^2 (M+m)^2}{5.32Mm}\right]^{1/2}. \tag{113}$$

This reduces to the equipartition value $U_t = \Omega$ for very weak fields $E = 0$, while in strong fields the terminal energy becomes nearly proportional to the field according to

$$U_t = \lambda_0 E(M+m)/2.31Mm.$$

Since the root mean square speed is given by

$$C_1 = (2eU_t/m)^{1/2}, \tag{114}$$

we substitute Eq. (113) into Eq. (114), and this value of C_1 into Eq. (107) to obtain the general mobility equation

$$\mu = \frac{0.921e\lambda_0}{(2em)^{1/2}}\left[\frac{\Omega}{2} + \left(\frac{\Omega^2}{4} + \frac{\lambda_0^2 E^2 (M+m)^2}{5.32Mm}\right)^{1/2}\right]^{-1/2}$$
$$\times\left[1+\frac{m}{M}\frac{1}{\frac{1}{2}+\left(\frac{1}{4}+\frac{\lambda_0^2 E^2 (M+m)^2}{5.32Mm}\right)^{1/2}}\right]^{1/2}. \tag{115}$$

In this, the first term in brackets is a generalization of an analogous expression previously derived by one of us.[237] The second term in brackets replaces the factor $(1+m/M)^{1/2}$ which has appeared in previous mobility equations (*e.g.* Langevin[232]). Apparently in previous equations the effect on the mean free paths of the fact that the mean ionic energy exceeds the mean molecular energy (see discussion of Eq. (78)) has been neglected in this term,

which is of the nature of a correction term for persistence of velocities. Note that λ_0 is defined by $\lambda_0 = (\pi\sigma_{12}^2 N)^{-1}$ and is thus independent of velocities.

Eq. (115) reduces, for the two cases of chief interest, to: for electrons $m \ll M$, $\lambda_0 = \lambda =$ electron mean free path,

$$\mu_{\text{electrons}} = \frac{0.921 e\lambda_0}{(2em)^{1/2}} \left[\frac{\Omega}{2} + \left(\frac{\Omega^2}{4} + \frac{\lambda_0^2 E^2 M}{5.32m} \right)^{1/2} \right]^{-1/2} \tag{116}$$

for ions $m = M$, $\lambda_0 = \sqrt{2}\lambda$ where $\lambda =$ molecular mean free path,

$$\mu_{\text{ions}} = \frac{0.921 e\lambda_0}{(2eM)^{1/2}} \left[\frac{\Omega}{2} + \left(\frac{\Omega^2}{4} + \frac{\lambda_0^2 E^2}{1.33} \right)^{1/2} \right]^{-1/2} \times \left[1 + \frac{1}{\frac{1}{2} + \left(\frac{1}{4} + \frac{\lambda_0^2 E^2}{1.33\Omega} \right)^{1/2}} \right]^{1/2}. \tag{117}$$

Both equations reduce to simple forms in the limiting cases of very small field and very large field.

The *effect of inelastic impacts* is to increase the mobility. This result is not paradoxical when one notes that inelastic impacts diminish the translational speed of the particles, and that this speed occurs in the denominator of the mobility equation (107). If all collisions made by electrons were completely inelastic, their mobility is shown to be[237]

$$\mu_i = \left(\frac{\pi\lambda e}{2mE} \right)^{1/2} \tag{118}$$

which is at least in qualitative agreement with observation.[238]

(2) **Comparison of these applications of classical kinetic theory with experiment** can be made at several points.

(a) *The Maxwell distribution* of velocities is found among electrons thermionically emitted from a hot body (as already discussed under B (1)). Also in many parts of discharge tubes the electrons are found[239] (by methods to be described in Part II) to have a Maxwellian distribution with a mean energy determined jointly by the ionization potential of the gas and by the discharge conditions, rarely exceeding one-third of the minimum ionizing potential. Frequently, in regions of intense ionization near the cathode, there are found two superimposed Maxwellian distributions, one of relatively high mean energy apparently consisting of primary electrons from the cathode which have been scattered and the other of much lower energy comprising secondary electrons which have been produced by ionization of the gas. The mechanism by which these Maxwellian distributions are so quickly established in an ionized gas is not understood.* It is well known that electrons drifting through a gas under the influence of a

[238] Compton, Phys. Rev. **22**, 333, 432 (1923).

[239] Langmuir and Mott-Smith, G. E. Rev. **27**, 449, 538, 616, 762, 810 (1924).

* See, however, section D (4).

uniform field, and subject to the laws of conservation of momentum, acquire a terminal speed in which they have a Maxwellian distribution of velocities with respect to a system of coordinates moving with their uniform rate of drift,[240] but calculation of the number of collisions required approximately to attain this condition shows that in an ionized gas some additional and more effective agency must be chiefly responsible for the scattering of velocities. Generally speaking, electrons are found to have very close to Maxwellian velocity distributions in all parts of discharges where there is considerable luminosity, and also in the Faraday dark space. Wide departures from

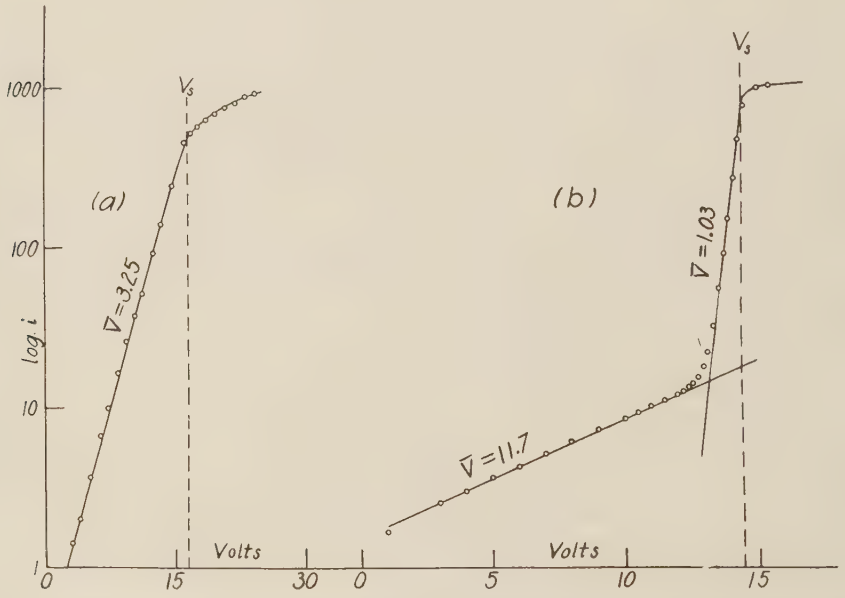


Fig. 21 (a) Showing one Maxwellian distribution. (b) Showing two superimposed Maxwellian distributions.

complete Maxwellian velocities are sometimes found in the dark spaces between striations, and always in the regions of cathode and anode falls of potential.

Fig. 21 shows two typical velocity distributions. The criterion of a Maxwellian distribution is that the graph of log i_- (electron current to a collecting electrode) against V (potential of collector) shall be a straight line in the region where V is negative with respect to space potential V_s. Its slope is $3/(2\overline{V})$ where $\overline{V}$ is the mean energy, in equivalent volts, of the electrons in the surrounding space. Fig. 21 (a) was obtained from an arc in argon, and 21 (b) from an arc in helium, both provided with incandescent cathodes as in tungar rectifiers.

[240] Langevin, Ann. Chim. Phys. **105**, 245 (1905); Boltzmann, "Gastheorie," Vol I, p. 114; Benade and Compton, Phys. Rev. **11**, 196 (1918).

Boltzmann's distribution of concentration is also found to hold for electrons in discharge tubes. A very pretty example of this will be found in a paper by Killian.[241]

(b) In the case of *mean free paths*, the experimental results point to a collision mechanism much more complicated than that assumed in the so-called "classical" kinetic theory case, discussed above; yet the results are in general of the same order of magnitude so that the classical kinetic theory values are useful for easy rough estimates or for use where no more accurate data are available or can be mathematically handled. There are three direct methods of measuring free paths of charged particles, due respectively to Lenard, Ramsauer, and Langmuir and Jones.

In *Lenard's* method[242] the charged particles are accelerated by a field to the desired velocity, a well defined beam of them is passed through a system of pinhole diaphragms into a chamber in which they may collide with gas molecules, and those which pass through undeflected are caught by a suitable receiving electrode. The number N getting through a distance x without deflection is $N_0\epsilon^{-x/\lambda}$ or $N_0\epsilon^{-px/\lambda_1}$, if λ_1 is the mean free path at unit pressure. The ratio of the numbers N' and N'' getting through at two pressures p' and p'', or at two distances x' and x'', is $\epsilon^{-x(p'-p'')/\lambda_1}$ or $\epsilon^{-p(x'-x'')/\lambda_1}$ respectively, whence the mean free path λ_1 at unit pressure is readily calculated.

It is necessary to examine what the mean free paths determined in any experiment really signify, since in each case the meaning of the mean free path is defined by the apparatus used to determine it. In the Lenard apparatus, a free path is one in which the angular scattering is very small,—too small to cause the particle to miss the collecting electrode. Energy losses are not involved, except in so far as they are accompanied by deflection.

In Ramsauer's method[243] Fig. 22, the charged particles are set free photoelectrically or thermionically as from the filament F, and are accelerated to the desired speed by an electric field between F and the electrode C. The beam emerging from the slit S_c is bent by a magnetic field to pass through the circular slit system S_1, S_2, S_3, S_4, S_5 and is finally caught in the Faraday box B.

The apparatus thus selects out a group of velocities very narrowly defined by the magnetic field and the slit system. The currents to the box B are measured for a series of different gas pressures and, as before, the ratio of currents at any two pressures is given by $\epsilon^{-x(p'-p'')/\lambda_1}$, where x is the length

[241] Killian, Phys. Rev. **35**, (1930).

[242] Lenard, Ann. d. Physik **12**, 714 (1903); Mayer, *ibid.* **45**, 24 (1914); **64**, 451 (1921); Compton and Benade, Phys. Rev. **8**, 449 (1916); Maxwell, Proc. Nat. Acad. Sc. **12**, 509 (1926); Jones, Phys. Rev. **32**, 459 (1928).

[243] Ramsauer, Ann. d. Physik **64**, 513 (1921); **66**, 546 (1921); **72**, 345 (1923); Jahrb. d. Radioakt. **19**, 345 (1922); Brode, Proc. Roy. Soc. **125**, 134 (1929); Phys. Rev. **34**, 673 (1929); Beuthe, Ann. d. Physik **84**, 949 (1927) probably vitiated by impurities; Brüche, *ibid.* 5, 909 (1929); Jones, Phys. Rev. **32**, 459 (1928) compares Ramsauer's and Lenard's methods for electrons in Hg vapor; see particularly Ramsauer and Kollath, Ann. d. Physik **4**, 91 (1930) for latest knowledge of behavior at very low velocities.

of path from S_c to S_5. Brode[243] has pointed out certain necessary precautions and second order corrections.

In this method it will be noted that not only angular deflection but also energy loss even without deflection will cause a particle to be lost from the beam. For this reason, the absorption coefficient $1/\lambda_1$ measured by this method, is often called the *total absorption coefficient.*

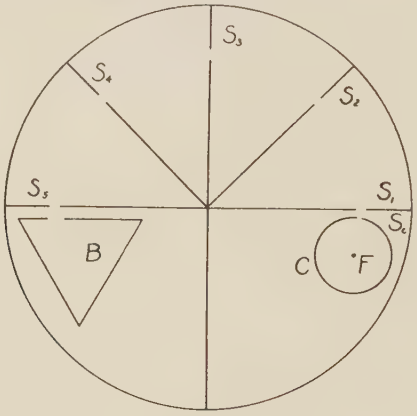


Fig. 22. Ramsauer's method for measuring mean free paths.

In *Langmuir and Jones' method,*[244] Fig. 23, ionization of the gas is maintained between an electron emitting equipotential hot filament F and anode plates A and B, while a surrounding coaxial cylinder C is used as a collecting electrode to which the current is measured at various potentials with respect to the filament. As will be shown in Part II, the main body of the enclosed

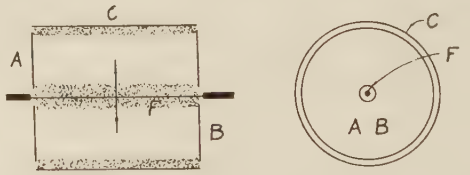


Fig. 23. Langmuir and Jones' method for studying electron collisions with gas molecules.

ionized gas is at a nearly uniform potential, approximately that of the anode, and differences of potential between this body of gas and the electrodes are concentrated in "space charge sheaths," around the electrodes, whose thicknesses are relatively small if the ionization is fairly intense. Thus the electrons emitted by the filament are projected out radially into the gas with velocities corresponding to the cathode fall of potential in the sheath. If the collector C is negative with respect to the filament F (and therefore with respect to all parts of the discharge) positive ions flow into it through the surrounding space charge sheath, but all electrons entering the sheath are

[244] Langmuir and Jones, Science, **59**, 380 (1924); Phys. Rev. **31**, 357 (1928).

forced back into the main body of the discharge, thus failing to reach the collector. But if the collector C is brought just to the potential of the filament, then those electrons from the filament which have lost no forward momentum are able to penetrate this sheath and reach the collector. For more positive potentials of the collector C, electrons which have lost some forward momentum are nevertheless able to reach it. Thus the current to this collector at its various potentials may be used to indicate the fraction of electrons leaving the filament which collide (lose forward momentum) while traversing the gas between filament and collector. Also this experiment yields much information regarding the nature of the collisions, as will be discussed in the following section. Fig. 24 shows a typical curve for nitrogen. In region AB only positive ions reach the collector C. At BC the cylinder is no longer negative in respect to F and the length $BC=\Delta$ is the relative change

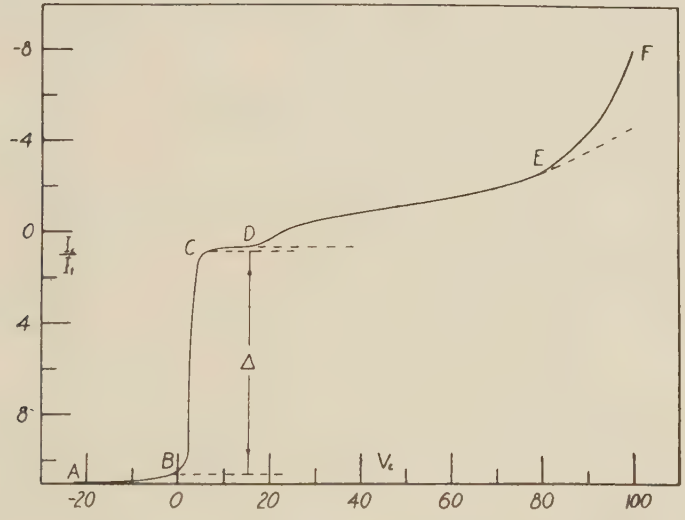


Fig. 24. Langmuir and Jones' results with nitrogen. Δ is the measure of the number of electrons which lose no forward momentum.

in current to the collector in passing from B to C. Then the fraction of primary electrons from the filament which reach the collector C without collision is

$$\frac{\Delta}{1-r+\beta_c} = \epsilon^{-ap/\lambda_1}$$

where r and β_c are small corrections due to reflection of a fraction r of the electrons incident on C and to change in the positive ion current to C due to the loss of this group of electrons as ultimate ionizing agents in the discharge, respectively, and a is the radius of the collector C. All these quantities may be measured or estimated with considerable certainty, whence λ_1 is found. Further information given by this method in regard to the nature of the collisions themselves is discussed in the following section.

Data of *electron mean free paths* from these three methods are shown in Figs. 25 and 26. Following common usage, the ordinates are "absorption coefficients" α of the gas at 1 mm pressure for electrons, or aggregate "effective" cross-sectional area of all the molecules in a cubic centimeter at 1 mm pressure, or average number of collisions made by an electron in 1 cm path through the gas at 1 mm pressure. These are all equivalent expressions and are equal to the reciprocal $1/\lambda_1$ of the mean free path at 1 mm pressure.

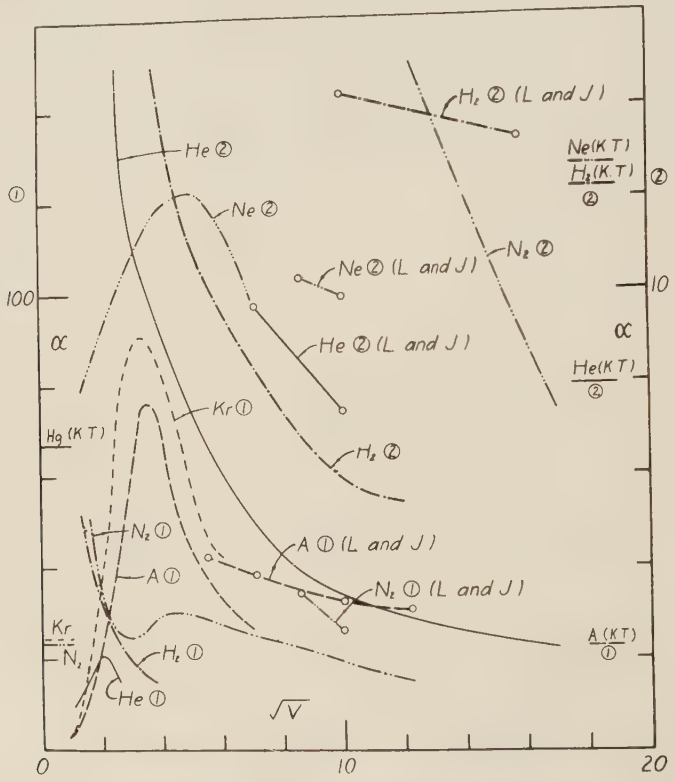


Fig. 25. Absorption coefficients $\alpha = 1/\lambda_1$ for electrons in gases at 1 mm, 25°C, as function of electron velocity. *K.T.* = Kinetic theory value from Eq. (82). Values by Ramsauer method. Values by method of Langmuir and Jones. (1) and (2) indicate scales of ordinates.

Data regarding *positive ion mean free paths* are very meager and are completely lacking in the most important case, that of positive ions moving among molecules of the same gas, except for protons moving through hydrogen.[245] Here Dempster found that protons of velocities between 50 and 2000 volts can pass through many hydrogen molecules without suffering appreciable velocity change, while Thomson found an angular distribution of scattering of the type which would be expected if the inverse cube law of force held during collisions. Similarly, for 900 volt protons in helium,

[245] Dempster, Proc. Nat. Acad. Sc. **11**, 552 (1925); G. P. Thomson, Phil. Mag. **1**, 961 (1926).

Dempster found abnormally long free paths [246] (or surprisingly small effects of collisions), while for protons of various velocities in helium and argon Thomson found maximum scattering (minimum free paths) at velocities of 1.4 $(10)^{-8}$ and 0.7 $(10)^{-8}$ cm/sec, respectively, in analogy with the "Ramsauer" effect for electrons.

More recently mean free paths of alkali ions in various gases have been studied. Kennard[247] and Thompson[248] have found that ions of lithium, sodium, potassium or caesium, moving in a beam through hydrogen, helium or argon, are much less scattered out of the beam than would be expected from consideration of the kinetic theory radii, and also that there is less

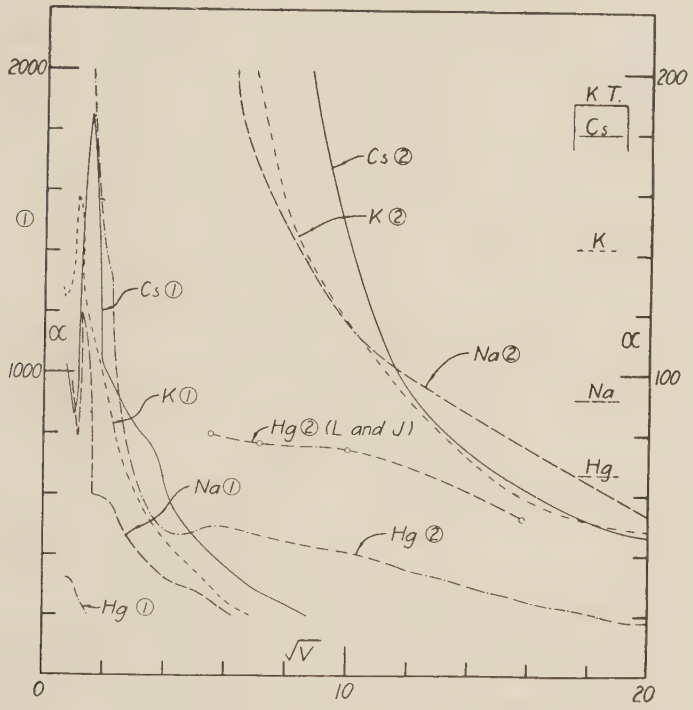


Fig. 26. Absorption coefficients α (see Fig. 25).

scattering at high than at low velocities. Ramsauer and Beeck[249] found also rapidly decreasing scattering at higher velocities, but found that the scattering approached a limiting value at velocities of the order of 50 volts or more, and that this limiting value agreed quite well with kinetic theory predictions. That is, the mean free path $\lambda_1 = 1/\alpha$ is such as would be expected if the effective collision radius is the sum of the radii of the gas molecule

[246] Dempster, Proc. Nat. Acad. Sc. **12,** 96 (1926).
[247] R. B. Kennard, Phys. Rev. **31,** 423 (1928).
[248] J. S. Thompson, Phys. Rev. (in print).
[249] Ramsauer and Beeck, Ann. d. Physik **87,** 1 (1928).

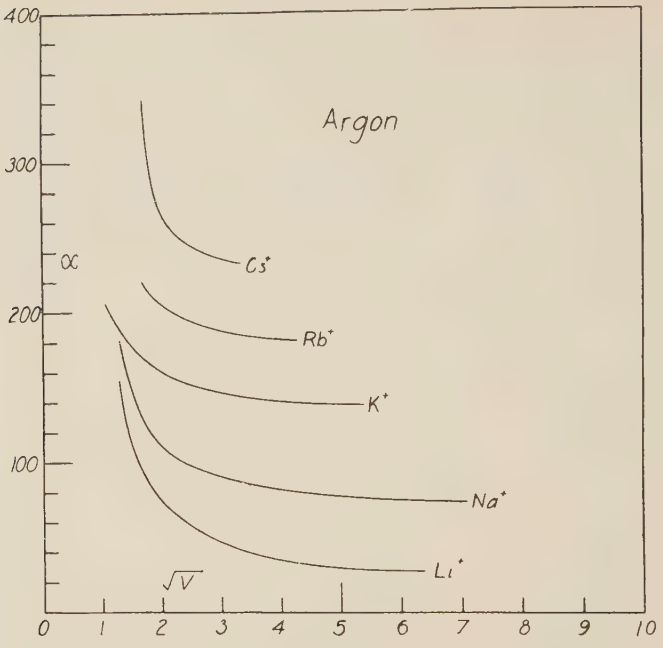


Fig. 27. Absorption coefficients $\alpha = 1/\lambda_1$ of alkali ions in argon gas.

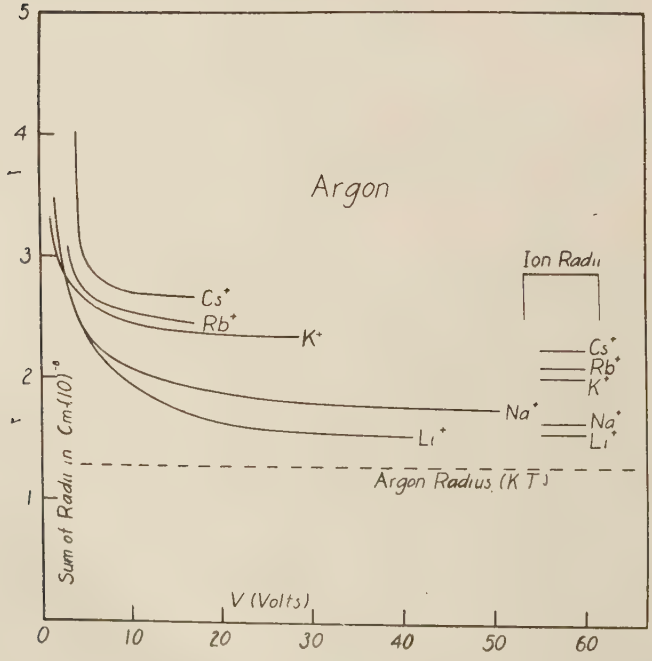


Fig. 28. Comparison of experimental with kinetic theory mean free paths of positive ions.

and the ion. (There is considerable diversity of evidence regarding the magnitudes of ion radii, as shown in the table on p. 22 of Ref. 249. The agreement is obtained if the "true sizes" of the ions are taken according to Fajans and Herzfeld.) Figs. 27 and 28 show a few of the results of Ramsauer and Beeck. In Fig. 28 the lines to the right of the curves and marked Cs^+, Rb^+, etc., indicate the sum of the radii of an argon atom and of each of these ions.

In order to reconcile the results of Thompson with those of Ramsauer and Beeck, it seems necessary to suppose that the mean free paths are actually of the order of those predicated by kinetic theory, especially at the higher velocities, but that the collisions result in loss of velocity without appreciable change in direction.

(c) *The nature of collisions and single scattering of electrons* by molecules has been investigated by two methods, the one devised by Dymond and the other by Langmuir and Jones. The former has the advantage of greater

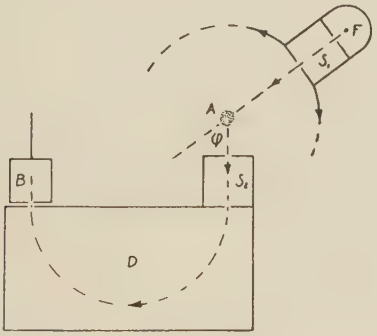


Fig. 29. Dymond's method of investigating electron scattering.

directness and resolving power, the latter of experimental simplicity. As will be seen, the two are in general agreement on the main features of electron scattering at collisions, though not in regard to some details.

In *Dymond's method*[250] electrons from a hot filament F (Fig. 29) are accelerated to any desired velocity v_1 and reduced to a well defined beam by a slit system S_1. Those which collide with molecules in a small region A and are deflected through an angle ϕ pass through the evacuated slit system S_2 into the evacuated deflection chamber D where those of any desired velocity v_2 are sorted out and deflected into the Faraday detection box F by means of a magnetic field (Dymond) or an electrostatic field (Harnwell). The scattering angle of the observations can be varied by rotating the electron "gun" FS_1 about the axis A. Thus the probability of deflection through any angle, and of any energy loss, can be found for

[250] Dymond, Phys. Rev. **29**, 433 (1926); Dymond and Watson, Proc. Roy. Soc. **122**, 571 (1929); Harnwell, Proc. Nat. Acad. Sc. **14**, 564 (1928); Phys. Rev. **33**, 459 (1929); **34**, 661 (1929); **35**, 285 (1930); Kollath, Ann. d. Physik **87**, 259 (1928); Whiddington, Phil. Mag. **6**, 889 (1928); Arnot, Proc. Roy. Soc. **125**, 660 (1929).

electrons of any initial energy colliding with molecules of any gas. Typical results of this method are shown in Fig. 30.

From a series of experiments such as those in Fig. 30, the angular distribution of scattering in helium is found to be as shown in Fig. 31 where, for comparison there are also shown some theoretical curves. The observations are seen to agree acceptably with the new quantum theory of scattering,[251] but *not* with classical kinetic theory. Furthermore the area under each particular curve gives the absorption coefficient α (reciprocal mean free path) for the corresponding type of collision. The most striking feature of these results is the relatively large probability of scattering through small angles, except at the very lowest velocities.

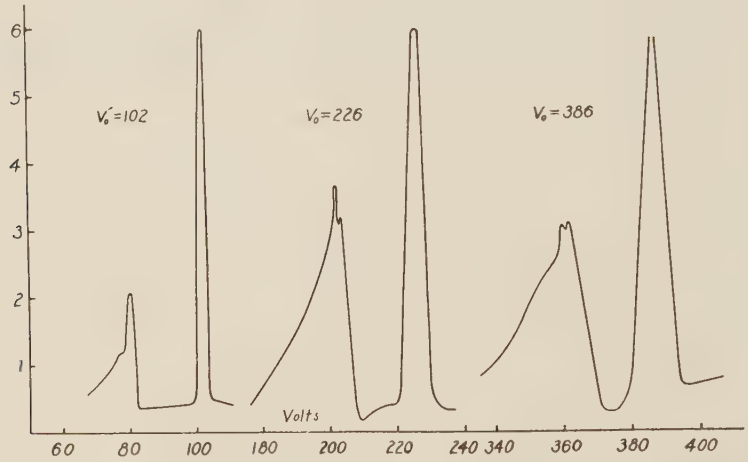


Fig. 30. Velocity distribution of electrons scattered at 10° by single impacts in helium, for initial velocities of 102, 226 and 336 volts. (Dymond and Watson [250].) The three high peaks represent elastic scattering; the pairs of lower peaks represent scattering with sufficient energy loss to excite one or the other of the two lowest excited states of the helium atom.

Harnwell's results[250] show that essentially the same characteristics are found in the scattering curves of helium, neon, atomic hydrogen, hydrogen and nitrogen. The average angle of elastic deflection appears to increase somewhat with increasing molecular weight.

In the Langmuir and Jones method[244] the apparatus of Fig. 23 was used to obtain data of the kind shown in Fig. 24. It will be recalled that the section AB of the latter figure represents positive ion current to the collector whose potential is V_c with respect to a homogeneous source of electrons which are projected at any desired speed directly toward the collector. When the collector changes from slightly negative to slightly positive with respect to

[251] For hydrogen, or in any Coulomb field, Born, Göttingen Nachrichten, p. 146 (1926); Zeits. f. Physik **38,** 803 (1926); Oppenheimer, Phys. Rev. **32,** 361 (1928); for scattering of particle or electron by a shielded nucleus, Sommerfeld, "Wellenmechanischer Ergänzungsband" p. 231. For scattering by helium, Mott, Proc. Camb. Phil. Soc. **25,** 304 (1929).

the source, those electrons Δ which have lost no forward momentum in the intervening distance become able to reach the collector. The information regarding the nature of collisions is obtained from a study of the shape of the current-voltage curve for positive values of V_c. The basis of interpretation of this curve is the following.

If all electrons which collided lost all their energy, it is obvious that the curve *CDEF* would be perfectly flat and would correspond just to those electrons which pass from source to collector without undergoing collisions. If, however, some electrons are slightly and elastically deflected, these

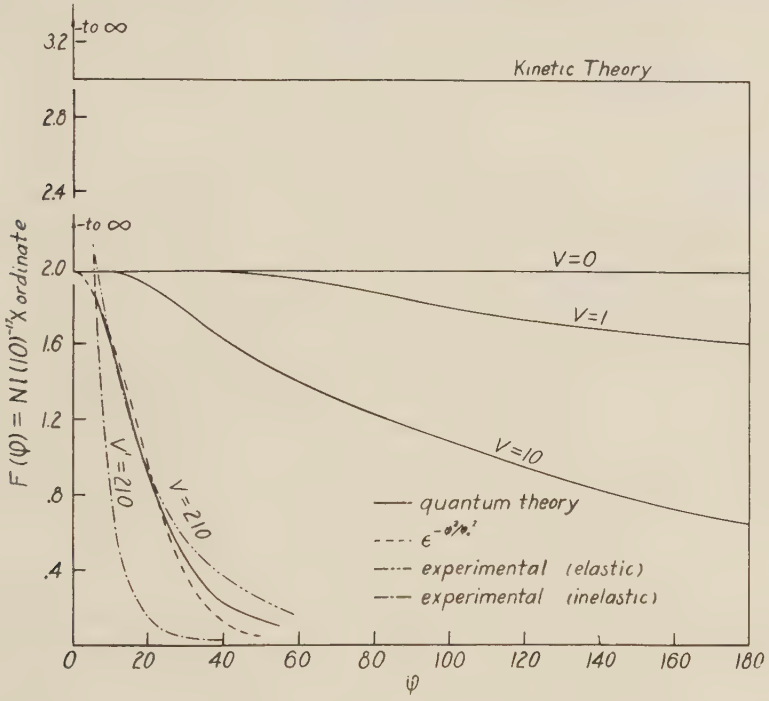


Fig. 31. Electron scattering in helium. $F(\phi)$ is the fraction scattered per unit solid angle at angle ϕ. Observations on *elastic* scattering of 210 volt electrons compared with quantum theory predictions. Also theoretical scattering at other voltages. Also observations on *inelastic* scattering of 210 volt electrons. *The observations are plotted to an arbitrary scale* which is the same for both elastic and inelastic scattering. Curve - - - is the graph of Eq. (119).

will not be able to penetrate the space charge sheath surrounding the collector when $V=0$, but they will be able to penetrate it when V is sufficiently positive to make the potential drop in this sheath less by an amount corresponding to the lost energy associated with the forward component of momentum. Thus the shape of the rising curve *CD* gives information in regard to the angular distribution of elastic scattering.

The break in the curve at *D* comes at about $V_c=13$ volts. It is at this voltage V_c that an electron, which has lost just enough energy to excite a nitrogen molecule and has continued to move directly forward, is just able

to penetrate the sheath and reach the collector. At still higher values of V_s electrons which have lost additional energy or been deflected in the process of exciting the atoms, can be collected.

In making the calculations, which are described in detail in the original paper, several secondary effects such as electron reflection at the collector and variation in total ionization as electrons are collected, must be taken into account. The experiment does not distinguish sharply between loss of energy and angular deflection, since loss of forward momentum is the quantity directly given. But some internal evidence, as well as supplementary experiments, justify the deduction of Table XVII.

The results are consistent with a probability law of angular scattering of the form

$$F(\phi) = F(\phi)_0 \epsilon^{-\phi^2/\phi_0^2} \tag{119}$$

where $F(\phi)$ is the fraction of the electrons which, in traversing unit distance, are scattered per unit solid angle at the angle ϕ, $F(\phi)_0$ is this fraction at the angle $\phi = 0$, and ϕ_0 may be called the *mean square angle of deflection*. Fig. 31 shows that such an equation is in approximate agreement also with Dymond and Watson's results in helium, the value of ϕ_0 there being 22.7°, for elastic scattering of 210 volt electrons. While this equation may not be exact, it is in fair accord with the facts and is of a form which can easily be handled in equations. It is suggested, therefore, that it may be quite useful in dealing with the influence of angular scattering in actual discharge devices. The appropriate constants to be used in this equation for several gases may be obtained from Table XVII, until more direct measurements are available as are now in the case of helium.

The total probability P of scattering as the result of a collision while going unit distance is given from Eq. (119) by

$$P = \int_0^\pi 2\pi \sin\phi \cdot F(\phi)_0 \epsilon^{-\phi^2/\phi_0^2} d\phi = \pi\phi_0^2 F(\phi)_0 \tag{120}$$

Thus, approximately,

$$F(\phi) = \frac{P}{\pi\phi_0^2} \epsilon^{-\phi^2/\phi_0^2}. \tag{121}$$

Values of P and ϕ_0 derived by substituting experimental results in this equation are given in Table XVII. From these, the value of $F(\phi)$ for any angle is calculable.

In some cases the curve above the first excitation potential was unsuitable for accurate estimates of ϕ_r. Since ϕ_r was found generally to be about equal to ϕ_e, it was assumed equal to ϕ_e in those more uncertain cases. These values are given in parentheses.

In applying Eq. (121) ϕ_0 is placed equal to ϕ_e or ϕ_r as the case may be, and the corresponding value of P is used.

TABLE XVII. *Summary of data on electron impacts in gases. (Langmuir and Jones.) P_k, P_e, P_r, P_i are the probabilities that an electron, while going 1 cm through gas at 1 mm pressure at 20°C, will collide inelastically, elastically, so as to excite the first excited state, or so as to ionize, respectively. ϕ_e and ϕ_r are the root mean square angles of deflection for elastic scattering and for scattering accompanying the first excitation respectively. $E_e = V\phi_e^2$ (volts radians²) is found to be "approximately" a constant for each gas and is useful for predicting ϕ_e for other electron energies V (volts). Note that ϕ_e must be expressed in "radians" in the argument of Eq. (121). $P_k + P_e$ = total probability of colliding.*

Gas	Volts	P_k	P_e	P_r	P_i	ϕ_e	ϕ_r	E_e
He	50	7.7	2.1	2.8	0.9	25°	(25)°	9.5
	100	6.0	1.5	0.7	1.6	19	16	11.0
Ne	75	9.3	1.1	1.5	1.8	21	(21)	9.8
	100	9.2	.9	1.3	2.4	19	(19)	11.0
A	30	19.1	24.1	5.9	4.7	24	(24)	5.3
	50	18.4	20.8	1.9	9.5	18	(18)	5.2
	100	18.7	14.5	1.7	11.4	12	(12)	4.7
	150	18.4	12.8	2.7	11.4	10	(10)	4.5
Hg	30	48.0	33.3	17.3	13.6	17	(17)	2.5
	50	49.1	29.5	14.7	19.9	11	13	3.1
	100	50.3	25.9	16.7	21.7	10	12	3.3
	250	32.0	21.3	6.7	20.4	6	(6)	3.2
H_2	100	9.2	5.3	3.5	3.9	5	16	7.3
	250	7.2	6.3	2.1	3.6	9	5	6.7
N_2	75	20.2	14.9	5.3	9.3	16	(16)	5.9
	100	16.2	10.4	4.8	10.3	14	8	5.8

Comparison of these results with those of Dymond and Harnwell described previously, shows agreement in the preponderance of scattering through small angles and in the magnitude of the mean scattering angle. Langmuir and Jones find inelastic collisions relatively more probable than do Dymond and Harnwell. In both cases the observed distribution $F(\phi)$ for small velocities tends to approach that of spherical particles.

In the important case of scattering in mercury vapor, Table XVIII gives data by both methods.

TABLE XVIII. *Single scattering of electrons in mercury vapor. $F_e(\phi)$ is the number of electrons elastically scattered through angle ϕ per unit solid angle per primary electron per cm path per 0.001 mm pressure at 20°C. $F_k(\phi)$ is the corresponding quantity for inelastic impacts. 82 volt primary electrons.*

	Dymond's Method (Arnot[250])		Langmuir and Jones[244]	
ϕ	$F_e(\phi)$	$F_k(\phi)$	$F_e(\phi)$	Ratio
0°	—	—	0.226	—
5	0.127	0.167	.184	1.45
10	.051	.063	.102	2.00
15	.024	.025	.038	1.58
20	.0144	.0111	.0099	0.69
25	.0089	.0061	.0012	0.15
30	.0056	.0036	.0003	0.05
40	.0028	.0019	—	—
50	.0015	.0011	—	—
60	.0005	.0004	—	—

In Table XVIII the first values of $F(\phi)$ are calculated from the data in Arnot's paper. The latter values are calculated from data in Table XVII and Eq. (121), taking $P_e = 0.0204$, $\phi_e = 11.3° = 0.198$ radians.

(d) *Diffusion* of electrons and ions has never been studied experimentally with sufficient accuracy to serve as any test of theories such as were developed in section D(1d). It is much easier experimentally to measure mobilities, and thence to calculate diffusion constants by Eq. (105).

(e) *Mobilities* of positive ions have never been studied under conditions in which the nature of the ions has been sufficiently well established to make the results of much use in testing an equation such as Eq. (117). In this direction, therefore, there is more point in combining theory and experiment to investigate the nature of the ions than in testing the theory itself.

Mobilities of electrons are known in many gases over quite a range of fields and pressures and permit a test of theories. Here the only theory which pretends to a rigid derivation and which does not contain arbitrary constants is that expressed by Eq. (116). Following a common usage, this equation may be written in terms of the so-called "mobility constant" K (which is really not a constant, since it depends on temperature) defined as the mobility under standard conditions of temperature and pressure by

$$K = (p/760)(273/T)\mu. \tag{122}$$

If μ is substituted from Eq. (116) and values of constants are introduced, this becomes (since $\lambda = \lambda_0$ for electrons)

$$K = \frac{271{,}000\lambda_1(273/T)^{1/2}}{[1+(1+1{,}106{,}000M\lambda_1{}^2(E/p)^2)^{1/2}]^{1/2}} \tag{123}$$

where λ_1 is the electron mean free path at $p = 1$ mm and 273°K, M is the molecular weight on the basis of $M_H = 1$ and E is the field in volts per cm. The numerical constants are slightly different from those previously published[252] and are believed better to express the theory.

The experimental results come surprisingly near to the predictions of this theory, when one considers the rather large departures of actual collision phenomena from simple kinetic theory and fact that the equation contains no arbitrarily adjustable constants. Figs. 32, 33 show comparison of theory with experiment in two cases; other cases are discussed by Compton.[252]

These and similar results show that the theory gives results of the right order of magnitude and about the right type of variation of K with (E/p) even as regards the peculiar horizontal intercept of the curve with the axis at $E/p = 0$. Thus the general idea underlying the theory is probably correct and its inaccuracies are due to its use of the much too simplified kinetic theory idea of a collision. Just how the theoretical results would be modified by introducing experimental values of free paths is not immediately obvious. For example, Fig. 25 shows that experimental determinations of electron

[252] Compton, Phys. Rev. **22**, 432 (1923).

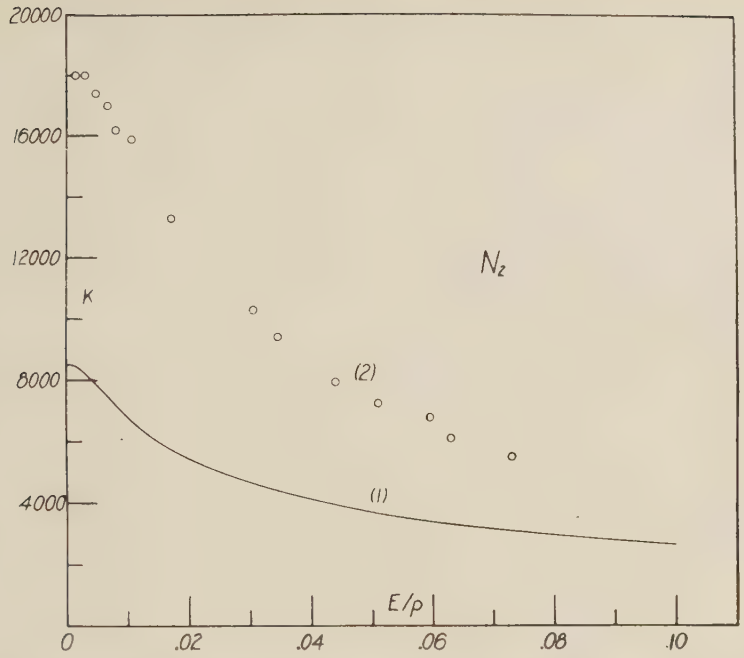


Fig. 32. Electron mobility "constant"; curve (1) theory by Eq. (123); curve (2) experimental results of Wahlin.[253]

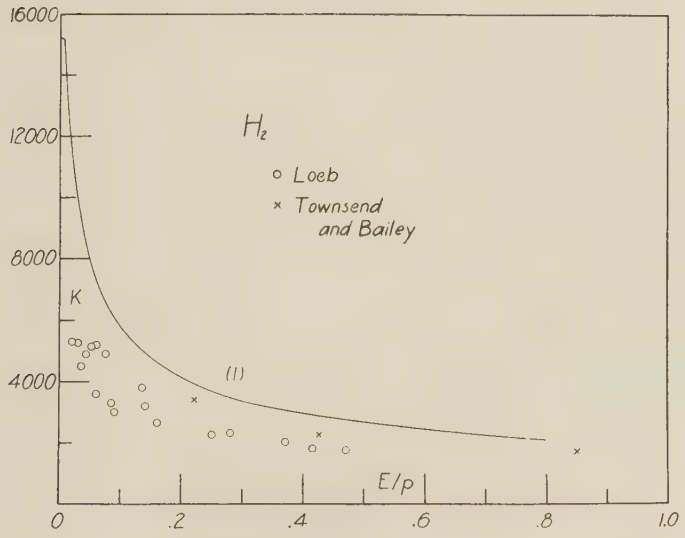


Fig. 33. Electron mobility "constant"; solid curve is theory by Eq. (123); experimental results by Loeb[254] and by Townsend and Bailey.[255]

[253] Wahlin, Phys. Rev. **23**, 169 (1924).
[254] Loeb, Phys. Rev. **19**, 24 (1922); **20**, 397 (1922) with small correction for wave form.
[255] Townsend and Bailey, Phil. Mag. **42**, 873 (1921).

mean free paths in N_2 give values considerably *smaller* than the kinetic theory value for such small velocities as are involved in mobility experiments. Introduction of such smaller free paths would, however, accentuate rather than remove the discrepancies shown in Fig. 32. Consideration of possible imperfect elasticity of impact will not bring agreement since it would not change the value of K for $E/p=0$, but would only modify the subsequent course of curve. In this case of nitrogen, however, the fact that scattered electrons are much more concentrated in the forward direction (small scattering angles) than would be true of elastic spheres would introduce a correction increasing the mobility and thus in the right direction to bring agreement between theory and experiment.

In hydrogen and helium, on the other hand, the experimental values of mobility constant are somewhat smaller than those predicted by the theory. In these cases, also, the experimental free paths at low velocities are much smaller than the kinetic theory values, which now is in the right direction to explain the discrepancy in Fig. 33, but the correction which would thus be introduced is somewhat too large, so that to obtain agreement it is again necessary to take account of the excess of scattering in the forward direction. In the case of hydrogen and helium, the quantum theory of scattering (which we have seen is well supported by experiments) predicts that this excess forward scattering is less pronounced at small than at large velocities and, for zero velocity, approaches uniformity of scattering through all angles, as in the simple kinetic theory case. It can fairly be said that such considerations seem adequate to modify Eq. (123) sufficiently to bring it into conformity with experiment, but no test of the theory along these lines has thus far been made.

Obviously a theory of mobilities based on the actual laws of electron scattering at low velocity impacts will be an extremely complicated thing. Perhaps, by assuming approximate analytical functions to represent the variation of free paths of and scattering angles with electron velocities, an approximate solution more satisfactory than Eq. (123) may be found.

There is another way of testing the mobility equations which avoids any consideration of terminal energy such as is involved in Eqs. (117, 123) and goes back to the more general Eq. (107), which involves only the assumptions that the distribution of velocities is Maxwellian and that all directions of motion are equally probable after a collision. The method of Langmuir and Mott-Smith already referred to[235] permits a test of the character of the velocity distribution and a determination of the average electron velocity in case the distribution is Maxwellian, provided the ionization of the gas is fairly intense. In such cases, Eq. (107) permits immediate calculation of the mobility μ if the mean free path is known, or vice versa. It is easy experimentally to determine μ since the total current carried by electrons equals $\int Ne\mu\, EdS$ integrated over the cross section of the current path, and N and E are given by exploring electrode measurements. Hence, in practice, the electron mean free path λ is calculated by applying the experimental results

in Eq. (107). In mercury vapor it is found[256] that the mean free path thus calculated is somewhat smaller than the kinetic theory value. This again is consistent with the experimental values shown in Fig. 26 for the small velocities here involved (for the most part of the order of 2 volts or less). Further tests by this method for a greater variety of conditions and for other gases are very desirable.

(3) **Quantum theories of collisions and scattering** are being reviewed (by E. U. Condon) in a separate article for this journal, and will therefore be treated here only in so far as is necessary to complete the discussion of these phenomena, which are so important in gas discharges. There are three cases which have been treated with considerable success by quantum theory: (a) *mean free paths for very small deflections*, such as would be involved in mean free path measurements; (b) *distribution of angular scattering as a function of velocity*, which gives a theory for the type of scattering studied by Dymond and Harnwell; (c) *abnormal transparency of some molecules for very slow electrons*,—the "*Ramsauer effect*," as observed apparently for all atoms or molecules with completed outer shells of 8 electrons. All of these apply only to electrons but (a) could be extended to include deflection of positive ions. We shall discuss these three cases very briefly.

(a) *Electron mean free paths for very small deflections* were investigated by Zwicky[257] by applying the method of perturbations after the well-known methods of the older quantum theory. The problem may be stated thus: "how far from the center of a molecule must the initial trajectory of an electron pass in order that it shall undergo a deflection ϕ?" Since the present theory applies only to small values of ϕ, the perturbation method is used as follows: First assume that the electron continues undeflected in the direction of its original trajectory. While passing, it exerts a distorting force on the molecule. Then the force exerted on the electron by this distorted molecule is used to calculate its deflection. This force includes that due to polarization $(K-1)e^2/(2\pi Nr^5)$ where K is the dielectric constant, N the number of molecules per unit volume and r the distance between electron and molecule. The force may also include a contribution from the permanent field of the molecule, if the latter is a permanent dipole, and if an electron in its orbit in the molecule has a period τ_1 which is comparable with the time τ_2 taken for the electron to pass the molecule, there will be a sort of resonance effect due to the behavior of the molecule as a dipole on account of the orbital electron, even though it would not act as a dipole in phenomena of much greater natural period. It is found that the deflection ϕ is proportional to

$$\phi \propto \frac{U(a)}{mv^2/2}\phi\left(\frac{\tau_1}{\tau_2}\right) \tag{124}$$

where $U(a)$ is the potential energy of the electron at the distance a, a is the closest distance of the molecule from the original trajectory of the electron, τ_2 is defined by $\tau_2 = a/v$, and ϕ is sort of "resonance function."

[256] Langmuir and Mott-Smith, G. E. Rev. **27**, 819 (1924); Killian, Phys. Rev. **35**, (1930).
[257] Zwicky, Phys. Zeits. **24**, 171 (1923); Proc. Nat. Acad. Sc. **12**, 461 (1926).

If the molecule possesses a high degree of symmetry the potential energy is due only to the polarization force above and

$$\phi \propto \frac{(K-1)e^2}{4\pi N a^4 m v^2} . \tag{125}$$

Experimentally we find the fraction of electrons deflected through an angle greater than some particular ϕ_0 defined by the construction of the apparatus. That is, we find the fraction whose trajectories come within a distance from a molecule. This fraction is $Nq = N\pi a^2 = \alpha = 1/\lambda$, where q is the effective cross section of a molecule and α (see Figs. 25, 26) is the aggregate effective cross section of the molecules in unit volume. Thus, from Eq. (125) we find

$$\alpha v = \frac{v}{\lambda} = \text{const } (K-1)^{1/2} . \tag{126}$$

In Figs. 25, 26, the curve for H_2 is seen to resemble such a hyperbolic relation, and that for Hg is somewhat similar. In fact it is found that a quantitative evaluation of the constant in Eq. (126) leads to calculated values of α which are within about 10% of the actual values in the cases of H_2, Hg, Zn, Cd.

If, on the other hand, the molecule is a dipole of moment μ oscillating or rotating because of electron orbital motion of period $\tau_1 = 1/\omega$, then Eq. (124) takes the form

$$\phi \propto \frac{\mu e}{a^2 m v^2} \phi\left(\frac{v}{\omega a}\right)$$

or

$$\alpha v^2 = \text{const} \left[\phi\left(\frac{v}{\omega a}\right)\right]^{1/2} . \tag{127}$$

Here α should decrease with increasing v more rapidly than in the case of simply polarizable molecules, it should be larger than it would be if polarization alone were effective, it should increase with atomic volume for atoms in the same column of the periodic table, and it should reach a maximum value in the neighborhood of $v = \omega a$. All these characteristics are exhibited markedly by the rare gas and the alkali atoms and to a less degree by Hg and N_2.

It would seem, therefore, that this theory of Zwicky is based on those physical phenomena which are actually effective in producing electron scattering at collisions. The same considerations could be used in a theory of positive ion scattering through small angles, except that here mutual polarization would have to be considered.

(b) The *distribution of angular scattering of electrons as a function of velocity* has been treated by wave-mechanical methods for atomic hydrogen

and other one-electron systems by Born[258] and Sommerfeld[259] and by Mott[260] for helium. Some results of this treatment have already been discussed in connection with Fig. 31. It would take us too far afield here to develop the wave mechanical theory of scattering, but the fundamental ideas and the results may be presented.

The electron of mass m and velocity v is treated as a plane wave of wave length $h/m\,v$ and amplitude ψ, the product of ψ by its conjugate $\bar{\psi}$ at any point being the measure of the average charge density, or the probability of the electron being at that point. The wave function satisfies the Schrödinger equation

$$\Delta\psi+\frac{8\pi^2m}{h^2}(E-V)\psi=0 \qquad (128)$$

where E is the total and V is the potential energy of the electron. If proper values of E and V are substituted in this equation, and it is solved for ψ as a function of the coordinates, then the distribution of ψ gives the distribution of scattering. In applying this method, the values of V characteristic of the fields surrounding various types of atoms, must be known.

(c) *The Ramsauer effect*, or the great transparency of certain gases to electrons of low speed, is again rather well explained by wave mechanics.[261] This phenomenon is observed for those gases like Ne, A, etc., whose molecules are electrically so symmetrical that their electric fields fall off extremely rapidly with distance. Thus the function V of Eq. (128) has appreciable values only within a very small distance of each molecule. The molecule is, therefore, an object of very small linear dimensions in so far as V is concerned. On the other hand, an electron of very small velocity v has a relatively long equivalent wave length $h/m\,v$. Hence this combination of a molecule which possesses very small external field, and a slow electron of long equivalent wave length, presents a case analogous to that of a small obstacle in the path of a train of waves of large wave length: the obstacle has relatively little ability to scatter the waves, which pass by it practically unaffected.

(4) **Plasma oscillations of electrons and ions** constitute a unique type among the numerous kinds of oscillations which can be obtained with discharge devices, in that they are entirely independent of the constants L, C, R of the circuit and in other respects. They were first discovered by Penning,[262] were suggested as the cause of the unexpectedly rapid electron scattering

[258] Born, Göttingen Nachrichten, p. 146 (1926).

[259] Sommerfeld, "Atombau u. Spektrallinien, Wellenmechanischer Ergänzungsband," p. 231.

[260] Mott, Proc. Camb. Phil. Soc. **25**, 304 (1929).

[261] Faxen and Holtsmark, Zeits. f. Physik **45**, 307 (1927); Holtsmark, Naturwiss. **16**, 614 (1928); *ibid.* **17**, 365 (1929); Zeits. f. Physik **48**, 231 (1928); *ibid.* **55**, 437 (1929).

[262] Penning, Nature, August 28 (1926); Physica **6**, 241 (1926).

discovered in strongly ionized gases by Langmuir[263] and further studied by Dittmer,[264] and were partly explained by Langmuir and Tonks[265] as follows.

A "plasma" is defined as a region of ionized gas in which electron and positive ion concentrations are approximately equal. In such a region, if a group of electrons at x is displaced in the direction of x by an amount $\xi(x)$, subject to $\xi(x)=0$ at two parallel bounding planes, then there is created an electric force arising from the unbalanced space charges, which tends to bring these electrons back to the equilibrium condition of zero space charge. The change in electron concentration, and the resulting field given by Poisson's equation are,

$$\delta n = n\delta\xi/\delta x \quad \text{and} \quad dE/dx = 4\pi e\delta n$$

whence

$$dE/dx = 4\pi ne d\xi/dx$$

whose integral is

$$E = 4\pi ne\xi.$$

Since the restoring force Ee is thus proportional to the displacement ξ, it is obvious that the motion of each electron is simple harmonic according to the equation

$$m\ddot{\xi} + 4\pi ne^2\xi = 0,$$

whose solution gives the natural frequency

$$\nu_e = \left(\frac{ne^2}{\pi m}\right)^{1/2} = 8980n^{1/2}. \tag{129}$$

There is thus a natural oscillation frequency proportional to the square root of electron concentration n. An ordinary value of n in low pressure discharge tubes is $(10)^{10}$ cm^{-3}, which leads by Eq. (129) to $\nu_e = 9\ (10)^8$ cycles which corresponds to radio waves of about 33 cm wave-length.

The velocity of propagation of these waves is

$$v = \lambda\left(\frac{ne^2}{m\pi}\right)^{1/2}. \tag{130}$$

Since the velocity is proportional to the wave-length, the waves show high dispersion and the group velocity of the waves turns out to be zero. Thus, although the waves can propagate through space, they transmit no energy.

[263] Langmuir, Phys. Rev. **26,** 585 (1925).

[264] Dittmer, Phys. Rev. **28,** 507 (1906).

[265] Langmuir, Proc. Nat. Acad. Sc. **14,** 627 (1928); Tonks and Langmuir, Phys. Rev. **33,** 195 (1928).

The average field amplitude of these oscillations may be estimated by considering each volume element to be an independent harmonic oscillator in equipartition equilibrium with the electrons themselves, provided we know the lower limit of size of these volume elements. As a reasonable guess, we may assume it to be the cube of the "Debye" distance[266] given by

$$\lambda_D = \left(\frac{kT_e}{8\pi ne^2}\right)^{1/2} = 4.90\left(\frac{T_e}{n}\right)^{1/2} \text{cm} \tag{131}$$

where T_e is the "temperature" of the electrons. (λ_D is the distance at which the average potential near a charged plane in the ionized gas is $1/\epsilon$ of that of the plane.) With this assumption the total energy density of the electric field of the oscillations is

$$\frac{E^2}{8\pi} = \frac{3}{2}\left(\frac{1}{\lambda_D}\right)^3 kT_e$$

whence

$$E = 96^{1/2}\pi^{5/4}e^{3/2}n^{3/4}(kT_e)^{-1/4} \tag{132}$$
$$= 1.17(10)^{-6}n^{3/4}T_e^{-1/4} \text{ volts/cm}.$$

This relation probably gives too large a value of E since it is probable that the mimimum volume element should be larger than λ_D, though of this order of magnitude.

It is obvious that any homogeneous beam of electrons, in passing through an ionized gas, will tend to acquire a random distribution of velocities about the mean velocity on account of the fields of these plasma electron oscillations. These oscillations, therefore, provide a means of interaction between ions and electrons other than that arising from individual encounters or mutual forces.

The foregoing oscillations of the electron space charge with respect to the relatively inert positive ion space charge are the "plasma *electron* oscillations." There are also natural oscillations of the positive ions, which Langmuir terms "plasma *ion* oscillations." For this case a treatment along similar principles gives the frequency

$$\nu_p = \left(\frac{ne^2}{\pi M + ne^2 M\lambda^2/kT_e}\right)^{1/2} \tag{133}$$

Here, if the wave-length λ is small, the equation reduces to the form of Eq. (129), with the ionic mass M in place of the electronic mass m. For long waves, on the other hand, the frequency approaches $(kT_e/M)^{1/2}/\lambda$, which means waves travelling with velocity $(kT_e/M)^{1/2}$. These latter waves are analogous to sound waves travelling through the ionized gas and their

[266] Debye and Hückel, Phys. Zeits. **24**, 185, 305 (1923).

frequencies are usually lower than 5 $(10)^5$ per sec. The demarkation between these two types of plasma ion oscillations is roughly the Debye wave-length of Eq. (131).

In Part II of this article, the authors will discuss actual discharge types, such as arcs, sparks, glow discharges and coronas, and will attempt an interpretation of their characteristics in terms of the fundamental phenomena discussed in Part I.

In conclusion, the authors are glad to express their thanks to Mr. Edward S. Lamar for preparing the drawings and to Miss Anna E. Lahey for her assistance in preparing the manuscript.

Volume 2 *July, 1930* *Number 3*

REVIEWS OF MODERN PHYSICS

RELATIVITY AND ELECTRODYNAMICS

By W. F. G. Swann

Bartol Research Foundation

IN AN article of this length, it would be useless to attempt to deal in detail with the subjects covered by the title, and an apology may seem necessary for writing, in these days, a review of such time-worn fields. There are many well versed in electrodynamics and relativity, and experienced in the technique of their use, to whom this article will appear as no more than the most elementary and fragmentary of treatments. To these, its interest will be confined to a few variations in point of view here and there, and possibly as a reminder that more fundamental elements of philosophic interest lie behind some of the uses of these sciences than are sometimes realized by those who are able to appease their philosophic conscience in contemplation of the beauty of mathematical analysis. On the other hand, there are those who have not specialized in these fields, to whom electrodynamics has its clearest visualization in engineering problems, to whom there remain curious thoughts concerned with whether or not lines of force move, those to whom the application of electrodynamics to things like electrons assumes a more and more shadowy meaning until nothing but the ghost of a meaning remains in the subject as applied to such fields as that of wave mechanics. There will be those to whom relativity is a subject having to do with an attempt to correlate the Michelson and Morley experiment with common sense, and which by some mysterious means has flown on the wings of mathematical analysis to all domains of physical theory, nonchalantly hiding the story of how it got there, behind an imposing array of symbolism which they are challenged and required to dicipher before the answer to the story will be vouchsafed. It is mainly to all such as these that the present article is addressed; and, it is hoped that, while its perusal may still leave them with many problems of mathematical difficulty, it may remove from the horizon some matters of perplexity.

In any presentation of a branch of modern physics, two courses are open. The first is the historical. This has the disadvantage that, usually, it does not represent a sequence of logical developments. The ways in which conclusions are reached are founded frequently upon considerations of special

cases, and sometimes are based upon experiments whose representatives have completely evaporated in the more general fields in which the conclusions are subsequently used. The alternative method is to take the results which have been stumbled upon in the historical development, review the path by which they have been reached, remove from it as much as possible of the obsolete debris with which it is encumbered, and try to construct some more satisfactory path by which the results might have been reached even to the extent of possible modifications in the starting points. This method runs the danger of appearing artificial. Many of the proofs which seemed fundamental to the old structure now play no part at all, and the whole vision may be apt to appear as a constellation in the sky instead of something standing on the rocks of the firm earth. However, an honest and unbiased facing of the situation in such a case will usually reveal that, in the last analysis, the older and apparently substantial concepts will not bear transportation to the realms in which the subject claims some of its most powerful uses. Those things which seem to endow the subject with such elements of tangibility when considered on an engineering scale dissolve in meaning, leaving only the fundamental content which, as a matter of fact, was the thing really of moment in the engineering problems where, however, it remained hidden in the glittering robes of a spurious reality. And so, to the critically minded, there appears a reality in the new artificiality and an artificiality in the old reality. The process of generalization is often not one in which the new forms a complication, or extension of the old, but rather one in which the old is a rather vague, somewhat incomplete, and illogical application of the principles of the new.

In what follows, it will be our purpose to dissect as far as possible the real content of the formulae, laws and concepts dealt with. When, as the result of a merciless stripping of irrelevant adornment from the laws, we have a spectacle impalatable to the intellectual taste, let the reader ask himself whether this cold remnant does not contain the whole essence of the laws in the sense that they are actually used. Our thoughts are accompanied by many visions which play no part in their essential structure. While we would strenuously deny that our picture of an all-pervading ether was like that of water, many of us supplement the cold statement of properties of medium with vague shadows of substantiality concerning which only the most dire intellectual torments would bring us to confess, even to ourselves, that we were semi-consciously thinking of the taste of the ether, of its boiling point and freezing point, of its latent heat of evaporation and so forth.

The Theory of Relativity

The so-called restricted theory of relativity arose, as is well known, in connection with the interpretation of experiments designed to detect what, in the language of thought of the day, was the earth's absolute motion in space. We may conveniently dispense here with the history of the matter, pass over one or two stages in the journey of emancipation from these rather troublesome experiments regarded as a starting point, and state that, in the

hands of Einstein the content of the theory may be summed up in two statements as follows.

(1) If two observers A and B, moving relatively to each other with a velocity v, assign coordinates and time to some event recognizable by each of them, the values which they assign will not be the same, but will be such that if each of them uses his own measures to determine the velocity of the same set of particles shot out from a point, each will find for that velocity the same value, irrespective of the direction, provided that that value is a certain constant c associated with the velocity of light.

(2) The same relations between the above measurements of the two observers which serves to conceal from them any effect of their motion in the matter of their measurements of the speed of light, will also serve to conceal such effects as regards any tests in which they perform corresponding experiments in their respective systems.

These statements as they stand, are fraught with danger as to misinterpretation. No. 2 particularly is open to all sorts of misrepresentations, so it will be well to start by dissecting the statements in detail. Before doing this, however, I cannot resist making the remark that as one absorbs the philosophy of the theory of relativity, he soon comes to a point where he wonders what place these imaginary observers have in the scheme of things, and why the theory of relativity should have to do apparently so fundamentally with the velocity of light. In his suspicion of the fundamentality of these intermediaries he is justified. To dispense with them at this state, however, would be to depart too far from the historical development of the subject to give the mind of the reader the opportunity of attuning itself sufficiently closely to the habits of thought of those who have grown up with this development. He might fail to understand much that has been written on the subject or indeed to see any reason for its existence. For these reasons we shall defer the discussion of the more fundamental approach to the restricted theory of relativity, and shall for the present, follow the usual practise of talking of clocks and scales and the like.

Significance of Statement No. 1. Let us call our observers A and B, and let us suppose that each is supplied with scales by which he may measure distances, and clocks by which he may measure time; and suppose that the latter are distributed throughout all space, so that at any place at which the time is required for any event taking place, there is a clock conveniently by to record it. These clocks are simply to be regarded as symbolic; all we wish to suppose, in fact, is that there is some process by which each observer is able to arrive at what he calls the time of a recognizable event at any place. It may be partly by measurement and partly by calculation that he secures this end, as when an astronomer assigns the time of an eclipse allowing for the velocity of light in transit. We are not for the moment interested in any question of "correctness" of the measures. All we imply is that each observer has *some method* of assigning his times, and this fact is symbolized by his clocks. If B is moving relatively to A with velocity v, we are of course to

suppose that as viewed by A, B's clocks and scales will be found travelling through space with the velocity v.

Suppose now that from the origin of A's system of coordinates light particles are emitted in all directions with a velocity which A measures as c. How does A measure this velocity? We may suppose that he is provided with

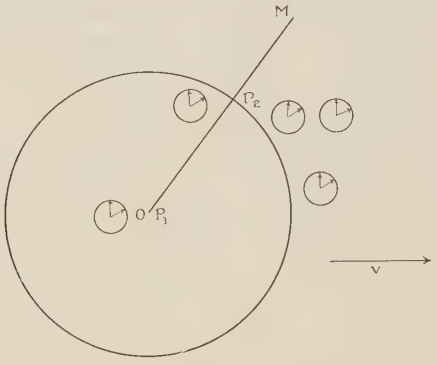


Fig. 1.

a scale OM, one of whose ends, of graduation P_1, is at his origin. He records the time t_1, as indicated by his clock at the origin when the particles are emitted. When a light particle passes some point reading P_2 on the scale the time recorded by one of his clocks located there is t_2. The amount $P_2 - P_1$

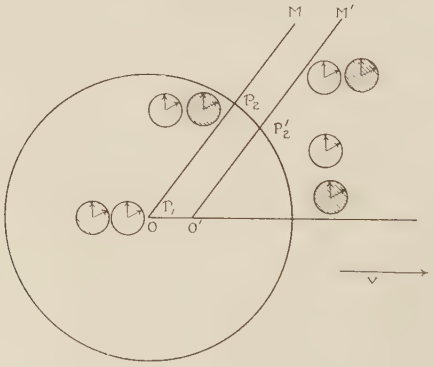


Fig. 2.

is then the value which he finds for the distance travelled by the light in the time $t_2 - t_1$ and we are supposing that the ratio is c *i.e.*

$$\frac{P_2 - P_1}{t_2 - t_1} = c.$$

Now let us superpose on Fig. 1, the apparatus owned by B and so obtain Fig. 2, where $O'M'$ is B's scale and the shaded clocks are B's clocks. Let us suppose that at the instant when the light signal was emitted from O the scales coincide, so that in the present picture, (Fig. 2) which is supposed to

correspond to the instant t_2 as measured by A, the scales will be shown in the positions indicated.

Let us start by supposing that the clocks and scales of A and B read alike in the sense that if any clock in B be found coinciding with a clock of A it will be found to have the same reading, and if any scale of B be found by B coinciding with a scale of A, it will be found also to coincide division for division with that of A if one pair of divisions coincide. Now the time recorded by B for the emission of the light particles is t_1, the same as that recorded by A since the scales coincided initially, and the initial reading on B's scale will be P_1. When the light particles are on the circle shown in Fig. 2, A will, as we have stated, record P_2 for the position of one of them on his scale, and t_2 for the corresponding time. B's scale will have moved to the position shown in Fig. 2, but the time by A's clock when the light particle reaches the point we have designated as P_2' will obviously be t_2: and if we assume B's clocks to read the same as those of A, the time recorded by B for this coincidence of the light particle with the mark P_2' will also be t_2. Thus the difference of the scale readings recorded by B for the light particles journey in the time t_2-t_1 is $P_2'-P_1$ and the velocity which he will calculate for the light particle is

$$c' = \frac{P_2' - P_1}{t_2 - t_1}.$$

Since we are supposing that one of A's scales if superposed on one of B's would coincide with it mark for mark, it is obvious that P_2' will be less than P_2, and hence the velocity c' measured by B will be different from the velocity c measured by A. If, however, B actually calculated from his measurements the same velocity c as is found by A, then either his scales or clocks or both record differently from those of A. Now Einstein set himself the problem of ascertaining how the readings of the scales and clocks of B must be related to those of A in order that both should find the same value c for the velocity of light. The answer to this question was as follows: Suppose one observer A adopts a rectangular system of coordinates designated x, y, z, and suppose t is the time which he records for a certain event at x, y, z. Then the corresponding values x', y', z', t' which the observer B will record for the event are such that

$$x' = \beta(x - vt) \tag{1}$$

$$y' = y \tag{2}$$

$$z' = z \tag{3}$$

$$t' = \beta\left(t - \frac{v}{c^2}x\right) \tag{4}$$

where

$$\beta^2 = \frac{1}{1 - v^2/c^2}$$

and where it is supposed that, in the frame of reference of A, B is moving along the axis of x with velocity v.

No attempt is made to *explain* why there should be any such difference in the measures. Thus, older, more intuitive attempts to correlate the experimental facts, attempts which preceded the theory of relativity, sought some physical reason as to why the scales and clocks of an observer should change in their readings when their velocity was changed. For the moment at any rate, we shall disregard all thoughts of this kind, with their not too immediately obvious pitfalls in logical thinking, and we shall concentrate our attention on the state of affairs depicted by (1)–(4) as our starting point.

The reader may complain that we have said nothing about the way in which B adjusts his scales and clocks to warrant the belief in any such conclusion as was reached in (1)–(4). That is true. It is customary to start by making a definition of simultaneity founded on the idea that if in either system an observer sends a light signal from a point "a" at t_1, reflects it at "b" back to "a" where it arrives at t_2, the clock at "b" will be running synchronously with that at "a" if the time t_3 which it records for the arrival of the light signal is $t_3 = \frac{1}{2}(t_1 + t_2)$. This provides a logical method for setting the clocks, a method which is the same for each observer, and one which can easily be seen to be self-consistent and to perpetuate its consistency if the relations (1)–(4) are true. Moreover, we can leave to the reader such trivialities as to how the two observers decide that one of them is not measuring in meters and the other in yards. I have wished to suppress this discussion about the setting of the clocks, and the like, however, because it is just one of those elements which has really nothing particularly to do with the vital elements of the theory of relativity. It is something necessitated to provide a comfortable existence for these imaginary observers who do not exist. Perhaps it will be well to refrain from confusing matters too much by discussing this point at present, so we shall leave it now, with a promise to return to it later when the time is more ripe.

Significance of Statement No. 2. Now a primitive view of the way in which nature acts might suggest that if we measured the resistance of a wire when its length was parallel to what we suppose to be the line of the earth's motion and then turned it through 90°, the resistance might be changed, whereas there would be no change if the earth were at rest as we understood that term. Experiments of this and various analogous kinds have been tried and all failed to give any indication of a motion for the earth in the sense sought. In all such experiments, the experimenter sets up certain apparatus —he assigns coordinates x, y, z, to certain things. Possibly he may also produce certain velocities or accelerations of certain parts, but as a result of all his manipulations he finally observes something —the motion of a galvanometer spot for example. If we are thinking of two observers (possibly with different measures of space, x, y, z, and time t), and if we are supposing them to be performing identical experiments, we are to understand by that, that each does the same thing in his own system of coordinates. If one places a

source of light at the point $x=2$, $y=4$, $z=6$, the other places a similar source of light at $x'=2$, $y'=4$, $z'=6$. If one arranges for the velocity of a particle to be $dx/dt=8$, $dy/dt=0$, $dz/dt=3$, the other arranges to have $dx'/dt'=8$, $dy'/dt'=0$, $dz'/dt'=3$. With this understanding as to the similarity of two experiments, two observers having relative motion will, on account of the Doppler effect observe different frequencies for a given spectrum line of a star. This fact constitutes no contradiction to the above statement, however; for here, the velocity of the star relative to each of the two observers is not the same. If each observer had his own star and if A's star moved relatively to him with the same velocity and direction as the other star moved relatively to B, the nature of the light observed by the two observers would be the same. Thus, one of the fundamental postulates of the restricted theory of relativity, in its historical aspect, at any rate, is that two observers moving relatively to each other with uniform velocity v obtain the same results in similar experiments. In such experiments as have been done, the hypothetical observers are at most really the same observer experimenting at two different times of the year when the direction of the earth's motion is not the same, and in the case of most of the experiments even this aspect is a generalization of the conclusions drawn for experiments performed at one time of the year,[1] but with the apparatus oriented first in one direction and then in a direction at 90° to the first direction.

It is not my intention to discuss here the comprehensiveness of the experimental evidence for the postulate, but simply to trace the consequence of the postulate. Accepting the postulate as true it implies that the laws of nature—the laws governing all phenomena—must be of the same form expressed in terms of x, y, z, t of equations (1)–(4) as in terms of x', y', z', t'. For, if this were not the case, it would, in general, be possible for our two hypothetical observers to obtain different results in similar experiments, similar in the sense above described.

It may perhaps be well to restate, in other words, the content of what has been written above. The laws of physics are expressed in terms of differential equations. Even an equation without any differential coefficients in it is a limiting case of a differential equation. It is a differential equation of zero order. Now when we proceed to seek the solution of any problem it is necessary not only to solve the differential equation, but also to specify the concrete values of certain things at $t=0$, for example. Thus, the Newtonian equations of dynamics are equations of the second order, and we must assign two things initially for each coordinate, the coordinate and the velocity (or momentum). If the equations had been of the third order, it would have been necessary to assign three things for each coordinate at $t=0$, the coordinate itself, the velocity, the acceleration; and so on. In general, we may say that the act of two observers doing the same experiment in two different systems is to be symbolized by their assigning the same numbers to the quantities which

[1] D. C. Miller, for example, has called attention to the fact that the original Michelson-Morley experiment is often quoted as having been performed at various times during the year, whereas it was actually performed over a very limited period.

the differential equations requires to have assigned in order to complete their solution. In fact, the act of doing the same experiment in the two systems is symbolized by the assignment of the same initial conditions in each.

It is manifest that if the differential equations for the two observers are the same, they will always get the same solutions for identical initial conditions. Or, in other words, they will always get the same *results* when they do the *same experiments*, understood in the above sense. Moreover it is not difficult to see, that, unless the differential equations are the same in the two systems,[2] the observers will not in general obtain the same results for the same initial conditions.

Let us enlarge upon the foregoing statement, and by way of illustration, suppose that the Newtonian law of gravitation were under test as to its fitness to conform to the requirements we have stated. It will suffice to consider the law as applied to two bodies a and b, in which case, it takes the form

$$m_a \frac{d^2x_a}{dt^2} = \frac{m_a m_b G(x_a - x_b)}{\{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2\}^{3/2}} \tag{5}$$

$$m_a \frac{d^2y_a}{dt^2} = \frac{m_a m_b G(y_a - y_b)}{\{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2\}^{3/2}} \tag{6}$$

$$m_a \frac{d^2z_a}{dt^2} = \frac{m_a m_b G(z_a - z_b)}{\{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2\}^{3/2}}. \tag{7}$$

Now by means of the relations (1)–(4) it would be possible to express x_a, x_b, y_a, $y_b \cdots$ etc. in terms of x_a', x_b', $\cdots$, etc, and t', and to express d^2x_a/dt^2 etc. in terms of derivatives of the dashed spacial coordinates with respect to the dashed time coordinate. When all this was done, so that the transformed (5)–(7) had no longer any undashed letters in them, it might happen that the three equations contained as their consequence[3]

$$m_a \frac{d^2x_a'}{dt'^2} = \frac{m_a m_b G(x_a' - x_b')}{\{(x_a' - x_b')^2 + (y_a' - y_b')^2 + (z_a' - z_b')^2\}^{3/2}} \tag{8}$$

$$m_a \frac{d^2y_a'}{dt'^2} = \frac{m_a m_b G(y_a' - y_b')}{\{(x_a' - x_b')^2 + (y_a' - y_b')^2 + (z_a' - z_b')^2\}^{3/2}} \tag{9}$$

$$m_a \frac{d^2z_a'}{dt'^2} = \frac{m_a m_b G(z_a' - z_b')}{\{(x_a' - x_b')^2 + (y_a' - y_b')^2 + (z_a' - z_b')^2\}^{3/2}} \tag{10}$$

[2] If we could imagine a situation where the differential equations were not even of the same order, the very meaning of "doing the same experiment in two systems" would start to evaporate. Thus, for example, if the equations in B's system were of the third order, while those in A's were of the second, we might *define* the doing in B of an experiment similar to that done in A as the assignment of the same initial quantities in B as in A; but, under these conditions there would be an ambiguity in the results obtained in B, depending upon the unspecified values of the additional initial condition necessitated by the higher order of the equations in B.

[3] We do not imply that (5) might be expected to go into (8), (6) into (9), or (7) into (10); but simply that these three equations which we get by the transformation might be shown to contain (8)–(10) as their combined consequence.

Now such would not be the case in the instance cited, but if it had been,[3] the equations would have remained invariant under the transformation given by (1)–(4), and they would be consistent with the restricted theory of relativity. If our two observers A and B each in his own coordinates observed not the same pair of planets, but different pairs; and if it so happened that, for example, the quantities

$$\frac{dx_a}{dt}, \frac{dy_a}{dt}, \frac{dz_a}{dt}, \frac{dx_b}{dt}, \frac{dy_b}{dt}, \frac{dz_b}{dt}$$

were found to have the values 3, 6, 10, 4, 8, 2, respectively; and if x_a, y_a, z_a, x_b, y_b, z_b, had the values 2, 9, 7, 4, 2, 8, at $t=10$, and if B observed that for his planets

$$\frac{dx_a'}{dt'}, \frac{dy_a'}{dt'}, \frac{dz_a'}{dt'}, \frac{dx_b'}{dt'}, \frac{dy_b'}{dt'}, \frac{dz_b'}{dt'}, \quad x_a', y_a', z_a', x_b', y_b', z_b',$$

had, respectively, the values 3, 6, 10, 4, 8, 2, 2, 9, 7, 4, 2, 8, at $t'=10$, then the equations (8)–(10) being of the same form as (5)–(7) would insure that the subsequent history of B's two planets would go on in the same manner in the dashed system as A's would go in in the undashed system.

Of course, it is not to be supposed that if A and B observe the *same* pair of planets they will see motions of the same form. If one of these planets is going around the other in a curve which to A appears a circle, it will not appear a circle to B. It has no right to. It is obeying the same general law for A and B, the law as expressed by (5)–(7) or (8)–(10), but the initial conditions—the assigned positions and velocities at $t=10$ (using the symbolism of our example above) in the undashed system of coordinates, are not the same as the assigned values at $t'=10$ in the dashed system of coordinates, for the same planet.

There is perhaps a little more importance to be attached to the foregoing statement than might be apparent at first sight; for, there is here involved the rather delicate question of what we mean by a law of nature. Thus it might so happen that our observer A found that each of his planets described a circle around the other. Yet if he stated his law in this form, completing his specification as regards the size of the circle, etc., he would not make a statement which was invariant under the transformation (1)–(4), even though equations (5)–(7) had reverted to the type (8)–(10) under the transformation. It might be perfectly true that, to A, the planets did describe the circles to which we have referred. You will say, of course, we had no right to expect the forms of the orbits to be invariant, it is the differential equations (5)–(7) which are the things whose invariance is under examination, and these are the things which express the laws of nature. This is quite true, but it is as well that we should formulate for our own clarity the statement of precisely what we are to mean by the laws of nature in a discussion of this kind. Thus if we should take the circular orbits which we supposed viewed by our observer A, and differentiate the coordinates once with respect to

the time we should arrive at differential equations—and why are these not capable of being regarded as an expression of the law of nature? The answer to the question is that there would be left in the equations so formed constants which were peculiar to the problem we were considering. In the totality of all the problems to which our equations would be supposed to apply we should have to recognize the necessity of assigning different values to these constants for each problem. The equations (5)–(7) have no constants of this type. For the totality of all of the problems to which we consider them as applying the constants G and m_a are the same. When our equations have been so far robbed of special features (usually by reducing them to a sufficiently high order) as to have in them no constants which are to be supposed different for different problems, they may be said to stand as expressions of *laws* in the same sense implied when the word "law" is used in the theory of relativity.

The invariance we have discussed so far in speaking of equations (5)–(7) is that pertaining to equations in which there occur, in addition to constants, no quantities other than the coordinates and times pertaining to certain identifiable points. We have occasion, however, to speak of invariance in the case of such equations as the electromagnetic equations, where quantities such as the electric and magnetic fields $\boldsymbol{E}$ and $\boldsymbol{H}$ occur. What has been written above as to the significance of corresponding experiments in the two systems still applies in the sense that to perform like experiments in the two systems we must assign in those systems the same numbers to the corresponding quantities whose numerical specification is necessary to provide a solution of the problem with the aid of the differential equations. There remains, however, something to be said as to the significance of the quantities concerned in relation to some fundamental set of definitions, when the quantities in question are electric and magnetic fields, potential differences and the like. Thus, when I say "Let a planet occupy the position $x=3$, $y=6$, $z=7$, at $t=2$ and let its velocities be $\dot{x}=5$, $\dot{y}=9$, $\dot{z}=11$," I say something immediately interpretable in terms of the coordinates whose transformation is expressed by (1)–(4). When, however, I say, "Consider an electron in an electromagnetic field, $E_x=5$, $E_y=7$, $E_z=10$, $H_x=5$, $H_y=2$, $H_z=12$," it is, in general, necessary to know the definitions of these quantities in terms of something ultimately related to a moving point (as when we define the fields in terms of the motion of an electron) before I can attach a meaning to the transformation.

Thus, to anticipate for a moment the section on electrodynamics, it will be recalled that the electromagnetic equations may be written

$$\frac{1}{c}\left(\rho \boldsymbol{u} + \frac{\partial \boldsymbol{E}}{\partial t}\right) = \text{curl } \boldsymbol{H} \tag{11}$$

$$\rho = \text{div } \boldsymbol{E} \tag{12}$$

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t} = \text{curl } \boldsymbol{E} \tag{13}$$

$$0 = \operatorname{div} \boldsymbol{H} \tag{14}$$

where ρ is the electric density, $\boldsymbol{u}$ its velocity, $\boldsymbol{E}$ and $\boldsymbol{H}$ the electric and magnetic fields, and c the same constant as occurs in (1)–(4).

Now it is an algebraical fact that if (11)–(14) be subjected to the transformation (1)–(4), they assume the form:

$$\frac{1}{c}\left(\rho' \boldsymbol{u}' + \frac{\partial \boldsymbol{E}'}{\partial t'}\right) = \operatorname{curl}' \boldsymbol{H}' \tag{15}$$

$$\rho' = \operatorname{div}' \boldsymbol{E}' \tag{16}$$

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}'}{\partial t'} = \operatorname{curl}' \boldsymbol{E}' \tag{17}$$

$$0 = \operatorname{div}' \boldsymbol{H}' \tag{18}$$

where by the dashed symbols we understand as follows:

$$(\boldsymbol{u}_x', \boldsymbol{u}_y', \boldsymbol{u}_z') = \left(\frac{\boldsymbol{u}_x - v}{(1 - \boldsymbol{u}_x v/c^2)}, \frac{\boldsymbol{u}_y}{\beta(1 - \boldsymbol{u}_x v/c^2)}, \frac{\boldsymbol{u}_z}{\beta(1 - \boldsymbol{u}_x v/c^2)}\right) \tag{19}$$

$$\boldsymbol{E}_x' = \boldsymbol{E}_x, \quad \boldsymbol{E}_y' = \beta\left(\boldsymbol{E}_y - \frac{v\boldsymbol{H}_z}{c}\right), \quad \boldsymbol{E}_z' = \beta\left(\boldsymbol{E}_z + \frac{v\boldsymbol{H}_x}{c}\right) \tag{20}$$

$$\boldsymbol{H}_x' = \boldsymbol{H}_x, \quad \boldsymbol{H}_y' = \beta\left(\boldsymbol{H}_y + \frac{v}{c}\boldsymbol{E}_z\right), \quad \boldsymbol{H}_z' = \beta\left(\boldsymbol{H}_z - \frac{v}{c}\boldsymbol{E}_x\right) \tag{21}$$

$$\rho' = \beta\rho(1 - \boldsymbol{u}_x v/c^2). \tag{22}$$

Before we can say that the equations (11)–(14) are "invariant" under the transformation, however, it is necessary to insure that $\boldsymbol{u}'$, ρ', $\boldsymbol{E}'$, $\boldsymbol{H}'$, are the quantities which the observer in the dashed system really would recognize as velocity, density, electric field and magnetic field respectively. As regards $\boldsymbol{u}'$, the desired conclusion follows very easily, since (1)–(4) enables us to compare velocities in the two systems. In order to say anything about $\boldsymbol{E}'$ and $\boldsymbol{H}'$ it is necessary to appeal to the definitions of these quantities. We shall deal with this matter more fully later; and, it will suffice now to say as follows:

If $\boldsymbol{E}$ and $\boldsymbol{H}$ are defined in terms of the motion of some identifiable entity such as an electron, it is possible to write down the meaning of $\boldsymbol{E}'_y$, for example, in terms of the motion of something in the dashed system. We can then transform this to the undashed system so that $\boldsymbol{E}_y'$ in (20) becomes expressed in terms of velocities, accelerations, etc. of an electron in the undashed system. The relation now assumes the form.

Something about the motion of an electron in terms of the undashed coordinates $= \beta\left(\boldsymbol{E}_y - \frac{v\boldsymbol{H}z}{c}\right)$.

But this is a statement about the motion of the electron in relation to the fields $\boldsymbol{E}$ and $\boldsymbol{H}$. If it is true, it provides for $\boldsymbol{E}_y'$ being the quantity which the dashed observer recognizes in terms of his own definitions, as the y component

of the electric field. If it is not true the significance of E_y' as an electric field and with it the meaning of the invariance, falls to the ground. Similar remarks apply, of course, to the other quantities E_x', E_z', H_x', etc.

The Mathematical Procedure for Securing Invariance. Now as we have already remarked, equations (5)–(7) would not turn out to be invariant under the transformation (1)–(4); and, if, without examining the vital elements controlling the situation, anyone should try to modify these to a form which was invariant (without departing appreciably from the numerical story which they tell, since that story is known to be approximately correct), he would probably invite a confusion of symbols sufficient to occupy his efforts for many a long day, and he would conclude that any practical efforts along this line would be hopeless. In order that we may present the possibilities in a more favorable light, we shall make a temporary digression from the theory of relativity, and consider the situation which we really have in prerelativity physics, where something of the same kind of difficulty is really existent, but has become smoothed over in our thinking because the way has been carefully prepared for us so that some of us may never have realized the difficulties which we might have encountered had things been a little different from what they are.

Let me recall the well known-equations

$$4\pi i_x = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} \tag{23}$$

$$4\pi i_y = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} \tag{24}$$

$$4\pi i_z = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \tag{25}$$

which relate the current density i to the magnetic field H in the case of a system of steady currents. Suppose that we should wish to transform these equations to a system of axes which were inclined to the axes in terms of which (23)–(25) are expressed. If i_x', i_y', i_z', H_x', H_y', H_z', represent the corresponding components of our quantities referred to the new axes, it is obviously possible to express i_x, i_y, i_z, $\partial H_z/\partial y$, $\partial H_y/\partial z$, $\cdots$ etc. in terms of quantities with dashed letters. Thus for example, if l_1, m_1, n_1, are the directional cosines of the new x axis with respect to the old axes, l_2, m_2, n_2 and l_3, m_3, n_3 the directional cosines of the new y and z axes, respectively, with respect to the old,

$$i_x = i_x'l_1 + i_y'l_2 + i_z'l_3 \quad \text{etc.}$$

When we had completed our transformation, we should be greatly surprised if it should fail to result that the equations which we finally obtained contained as their consequence

$$4\pi i_x' = \frac{\partial H_z'}{\partial y'} - \frac{\partial H_y'}{\partial z'} \tag{26}$$

$$4\pi i_y' = \frac{\partial H_x'}{\partial z'} - \frac{\partial H_z'}{\partial x'} \tag{27}$$

$$4\pi i_z' = \frac{\partial H_y'}{\partial x'} - \frac{\partial H_x}{\partial y'}. \tag{28}$$

For if this consequence did not result we should say that the laws of nature depend upon the directions in which we choose our axes in space. Such a situation is not on the face of it necessarily absurd, for it might have been a fact that the universe had been constructed in a sort of huge crystal, with different properties in different directions. In fact, in the last analysis, such a state of affairs would constitute exactly what we should imply by the general statement that the universe had been constructed on a crystalline basis of this kind. The universe might have been so constructed, and, in limiting ourselves to laws which *do* retain their form unchanged when the axes are rotated, we implicitly postulate that it has *not* been so constructed.

But why is it that the law (23)–(25) *does* remain unchanged under the transformation? It is by no means an obvious conclusion. Anyone unfamiliar with the process, who tries the transformation for the first time would probably get into a hopeless muddle. And, to show how near things came to being different in this respect from what they are, it will suffice to remark that, had the minus signs on the right hand side of equations (23)–(25) been plus signs, the invariance of form under transformation would not have followed. Yet, the equations would have appeared in every way as natural and symmetrical as before. Why is it that the equations we happen to have chosen in physics do obey this criterion of invariance? It is because the quantities which we have equated are components of what we call vectors. We know that if we want to transform the coordinates of a point x, y, z to an inclined set of axes in which the directional cosines of the old x, y, z axes with respect to the new axes, are, respectively, l_x, m_x, n_x, l_y, m_y, n_y, l_z, m_z, n_z, the equations of transformation are:

$$x' = xl_x + yl_y + zl_z \tag{29}$$

$$y' = xm_x + ym_y + zm_z \tag{30}$$

$$z' = xn_x + yn_y + zn_z \tag{31}$$

Suppose we have three numbers P, Q, R, which are associated with the three axes, x, y, z, and three numbers P', Q', R', which are associated with x', y', z', in the same manner as P, Q, R, are associated with x, y, z. Then if P', Q', R', are related to P, Q, R, by the transformation

$$P' = Pl_x + Ql_y + R_z \tag{32}$$

$$Q' = Pm_x + Qm_y + Rm_z \tag{33}$$

$$R' = Pn_x + Qn_y + Rn_z \tag{34}$$

the three quantities P, Q, R, may be said to correspond to the components of a vector. This definition is as we shall see, far more fundamental than the

statement that a vector is a directed quantity, which, of itself means nothing unless one is able to say what he means by a directed quantity. If one speaks of a velocity he may think there is no ambiguity in his meaning as to its direction. But suppose he speaks of an angular momentum. We know that an angular momentum may be regarded as a vector, its direction being perpendicular to the plane of rotation. Now this is the one direction in which there isn't any motion at all. What then gives it the characteristic of being the line to represent the vector?

The answer is as follows: let us consider the case of a particle of unit mass moving in a closed orbit in a plane, and consider three quantities $(v_y z - v_z y)$, $(v_z x - v_x z)$, $(v_x y - v_y x)$, which we ordinarily regard as the components of an angular momentum, but let us for the moment regard them simply as algebraical expressions. The notation is sufficiently obvious to render explanation unnecessary. Calling the above three quantities A, B, C, for shortness, it can readily be shown that they are respectively proportional to l, m, n, the directional cosines of a line which is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$. Moreover, it is easy to see that the factor of proportionality is $(A^2+B^2+C^2)^{\frac{1}{2}} = |\boldsymbol{v}|\,|\boldsymbol{r}| \sin \Theta$ where $|\boldsymbol{v}|$ is the resulting velocity, $|\boldsymbol{r}|$ is the distance of the unit mass from the origin and Θ is the angle between $\boldsymbol{v}$ and the radius vector $\boldsymbol{r}$. In other words $(A^2+B^2+C^2)^{\frac{1}{2}}$ is what we call the resultant angular momentum, M.

Hence, since quantities like A, B, C, are always equal to M times the directional cosines of a fixed line whatever the set of axes chosen, A, B, C transforms according to (29)–(31) which are the equations of transformation of a line. Hence A, B, C, form the components of a vector and the direction of that vector is that determined by l, m, n. In other words, it is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$. Moreover, since also the sum of the squares of A, B, and C gives the square of the angular momentum, A, B, and C are called the components of a quantity, the angular momentum, whose direction is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$.

It is only in such a sense as that outlined above that there is any meaning to the statements that the angular momentum is a vector whose components are A, B, and C, and whose direction is perpendicular to $\boldsymbol{v}$ and to the radius vector $\boldsymbol{r}$.

Thus, a vector is a set of three quantities which transform as the components of a line transform under a rotation of axes.

Suppose now that we have two vectors whose components are P, Q, R, and X, Y, Z, respectively, and suppose that

$$P = X \tag{35}$$

$$Q = Y \tag{36}$$

$$R = Z \tag{37}$$

then, because each of these sets of three quantities constitute a vector,

$$P' = Pl_x + Ql_y + Rl_z \tag{38}$$

$$Q' = Pm_x + Qm_y + Rm_z \tag{39}$$

$$R' = Pn_x + Qn_y + Rn_z \tag{40}$$

$$X' = Xl_x + Yl_y + Zl_z \tag{41}$$

$$Y' = Xm_x + Ym_y + Zm_z \tag{42}$$

$$Z' = Xn_x + Yn_y + Zn_z \tag{43}$$

Now, because $P = X$, $Q = Y$, and $R = Z$, equations (38) to (43) show us that

$$X' = P' \tag{44}$$

$$Y' = Q' \tag{45}$$

$$Z' = R'. \tag{46}$$

In other words, equations (35)–(37) have reverted to the same type under the transformation.

We see that all that was necessary for this invariance to follow was that P, Q, R, and X, Y, Z, should transform in the same way. It was not even necessary that that way should be the same as the way in which the axes transformed. As a matter of fact the result (44)–(46) would have followed even though the l's, m's, and n's had not been constant but had been functions of x, y, and z, so that they were no longer directional cosines. All that would have been necessary would have been that the l's, m's, and n's of (41)–(43) should be the same as the corresponding l's, m's, and n's of (38)–(40).

The spirit of the idea is even more general than this. For, suppose we have any number of quantities P, Q, R, S, J, and another set of quantities X, Y, Z, U, T, such that

$$P = X, Q = Y, R = Z, S = U, J = T. \tag{47}$$

We must suppose that each of these quantities is defined in some way in the x, y, z system of coordinates. Suppose we consider another system of coordinates x', y', z' related to the first set by

$$\left.\begin{aligned} x' &= f_1(x, y, z) \\ y' &= f_2(x, y, z) \\ z' &= f_3(x, y, z) \end{aligned}\right\} \tag{48}$$

and define quantities P', Q', R', S', J', etc., X', Y', Z', U', T', etc. in the x', y', z' system of coordinates in the same way that the undashed quantities are defined in the x, y, z system of coordinates, and suppose that, when this is done, it can be shown that P', Q', R', S', J' are related to P, Q, R, S, J in the same way as X', Y', Z', U', T' are related to X, Y, Z, U, T. Suppose in fact

$$\left.\begin{aligned} P' &= a_{11}P + a_{12}Q + a_{13}R + a_{14}S + a_{15}J \\ Q' &= a_{21}P + a_{22}Q + a_{23}R + a_{24}S + a_{25}J \\ R' &= a_{31}P + a_{32}Q + a_{33}R + a_{34}S + a_{35}J \\ S' &= a_{41}P + a_{42}Q + a_{43}R + a_{44}S + a_{45}J \\ J' &= a_{51}P + a_{52}Q + a_{53}R + a_{54}S + a_{55}J \end{aligned}\right\} \tag{49}$$

and

$$\left.\begin{aligned} X' &= a_{11}X + a_{12}Y + a_{13}Z + a_{14}U + a_{15}T \\ Y' &= a_{21}X + a_{22}Y + a_{23}Z + a_{24}U + a_{25}T \\ Z' &= a_{31}X + a_{32}Y + a_{33}Z + a_{34}U + a_{35}T \\ U' &= a_{41}X + a_{42}Y + a_{43}Z + a_{44}U + a_{45}T \\ T' &= a_{51}X + a_{52}Y + a_{53}Z + a_{54}U + a_{55}T \end{aligned}\right\} \tag{50}$$

Then it is obvious that because of (47) we shall have

$$P' = X', Q' = Y', R' = Z', S' = U', J' = T'. \tag{51}$$

In other words the equation (47) will remain invariant under the transformation. It is immediately obvious that the result is not limited to three variables x, y, z. We might have had any number of variables $x_1, x_2, x_3, x_4, \cdots$ etc., so that (48) took the form

$$\left.\begin{aligned} x_1' &= f_1(x_1, x_2, x_3, x_4, \cdots) \\ x_2' &= f_2(x_1, x_2, x_3, x_4, \cdots) \\ x_3' &= f_3(x_1, x_2, x_3, x_4, \cdots) \\ x_4' &= f_4(x_1, x_2, x_3, x_4, \cdots) \end{aligned}\right\} \tag{52}$$

etc.

and the a's of (49) and (50) could have been any functions of x_1, x_2, x_3, x_4, etc.

Already we have passed beyond the stage necessary for the mere purpose of the restricted theory of relativity, but the more general conclusions we have reached will find their place in the general theory. The net result of our conclusions is that if two sets of quantities $P, Q, R \cdots$ etc.; $X, Y, Z, \cdots$ etc. transform in the same way, under a certain transformation of the coordinates a statement of equality between the two sets will be true in the transformed system of coordinates. As a special case of all this we have that where there are as many of the quantities $P, Q, R, \cdots$ etc., as there are variables $x_1, x_2, x_3, \cdots$ etc., then, of course, it is certainly sufficient that $P, Q, R, \cdots$ etc. shall transform in the same way as $x_1, x_2, x_3, \cdots$ etc. respectively, and that $X, Y, Z, \cdots$ etc., shall transform in the same way as $x_1, x_2, x_3, \cdots$ etc., in order that such relations as $P = X$, $Q = Y$, $\cdots$ etc., shall be invariant.

Now, we have seen that in the case of the Einstein transformation $x' = \beta(x - vt)$, $y' = y$, $z' = z$, $t' = \beta(t - vx/c^2)$.

As a matter of fact, it is customary for reasons of symmetry to change from the variable t to $u = ict$ where $i = \sqrt{-1}$, so that

$$x' = \beta\left(x + i\frac{v}{c}u\right), \quad y' = y, \quad z' = z, \quad u' = \beta\left(u - \frac{ivx}{c}\right). \tag{53}$$

Any set of quantities X, Y, Z, U, which transforms according to the law

$$X' = \beta\left(X + \frac{ivU}{c}\right), \quad Y' = Y, \quad Z' = Z, \quad U' = \beta\left(U - \frac{ivX}{c}\right)$$

constitute what is called a 4-vector. It is obvious from what we have said above that any equations between two 4-vectors will remain invariant under the restricted theory of relativity.

The restricted theory also makes use of sets of six quantities which transform in a certain way which is characteristic of all such sets. Such sets of six quantities are called 6-vectors. We shall not find it necessary to speak of them here. Thus provided that we decide to express our laws of physics as relations between 4-vectors, we shall quarantee our invariance from the start. But how are we going to be lucky enough to find such suitable 4-vectors in terms of which to express our laws? What was the corresponding problem in classical physics? How was it that we succeeded in talking in terms of vectors? It was because we *defined* the quantities which formed our fundamental concepts in such a way that they should be vectors. Thus the three components of a length, x, y, z constitute a vector (the most fundamental vector since it is the thing in terms of which the vector was defined). A velocity is a vector because it is simply a length divided by a time. An acceleration is a vector because it is a velocity divided by a time. A force is a vector because it is defined as proportional to an acceleration. To say that a force is a vector without specifying the sense in which it is defined is to talk nonsense.

And so, it turns out in the restricted theory of relativity that the reason we encounter such things as 4-vectors, naturally, is that we define the fundamental things which form the basis of our discussion in such a way that they shall be 4-vectors, or of course 6-vectors or scalars.

Thus, to take an illustration: if we consider any particle at any time t x_1, y_1, z_1, ict_1 constitute a 4-vector. If we consider the same particle, or any other particle at another place, x_2, y_2, z_2, ict_2 constitute a 4-vector, and since it is easy to see that the difference of two 4-vectors constitute a 4-vector, x_1-x_2, y_1-y_2, z_1-z_2, $ic(t_1-t_2)$. This is true even if the two particles are quite unrelated, just as those quantities obtained by taking the difference of the coordinates of two quite unrelated points in three dimensions constitute an ordinary vector. It is, of course, also true as a special case, that if x_1-x_2, y_1-y_2, z_1-z_2, represent the distances moved by a particle in the time t_1-t_2, then x_1-x_2, y_1-y_2, z_1-z_2, $ic(t_1-t_2)$ constitute a 4-vector. In the case where the differences are infinitesimal dx, dy, dz, $icdt$ still constitute a 4-vector. Now it is easy to show by making the substitution $dx'=\beta(dx-vdt)$, $dy'=dy$, $dz'=dz$, $dt'=\beta(dt-(v/c^2)dx)$ that

$$dx'^2+dy'^2+dz'^2-c^2\,dt'^2=dx^2+dy^2+dz^2-c^2dt^2.$$

In other words,

$$dx^2+dy^2+dz^2-c^2dt^2 \text{ is an invariant,}$$

so that

$$dt\left[1-\frac{(\dot{x}^2+\dot{y}^2+\dot{z}^2)}{c^2}\right]^{1/2} \text{ is an invariant.}$$

Calling $\dot{x}^2+\dot{y}^2+\dot{z}^2=w^2$, we have that $dt(1-w^2/c^2)^{1/2}$ is an invariant. It is customary to write $k\equiv(1-w^2/c^2)^{-\frac{1}{2}}$, so that k/dt is an invariant. Now,

multiplying the components of a 4-vector by an invariant, yields, of course, another invariant, so that

$$k\frac{dx}{dt},\ k\frac{dy}{dt},\ k\frac{dz}{dt},\ kic \text{ is a 4-vector,}$$

i.e., kw_x, kw_y, kw_z, kic is a 4-vector. Since k is very nearly equal to unity unless w is comparable with c, the first three components of this 4-vector are very nearly equal to the x, y, and z components of the velocity. The whole 4-vector is called the velocity 4-vector. Again, by taking the difference in the values of the velocity 4-vector at successive instants, we have $d(kw_x)$, $d(kw_y)$, $d(kw_z)$ $d(kic)$, is a 4-vector. Multiplying by the invariant k/dt, we obtain

$$k\frac{d}{dt}(kw_x),\ k\frac{d}{dt}(kw_y),\ k\frac{d}{dt}(kw_z),\ k\frac{d}{dt}(kic)$$

constituting a 4-vector. Since k is very nearly equal to unity, unless w is comparable with c, the first three components of this 4-vector are very nearly the same thing as the components of the acceleration. However, they are quantities well fitted to the use of the theory of relativity, because they are components of a 4-vector. Thus, if there should be other particles present, and we should build up some other 4-vector, P_x, P_y, P_z, icP_t depending upon the position of this first particle in relation to the others, and possibly upon the velocities of the particles, and if we should write down the relations

$$k\frac{d}{dt}(kw_x) = P_x$$

$$k\frac{d}{dt}(kw_y) = P_y$$

$$k\frac{d}{dt}(kw_z) = P_z$$

$$k\frac{d}{dt}(kic) = icP_t$$

the results might be wrong, but they would at least be invariant under the transformation of the restricted theory of relativity.[4]

[4] If, in ordinary vector algebra, we equate two vectors A_x, A_y, A_z, and B_x, B_y, B_z, it is necessary for consistency that $A_x B_x + A_y B_y + A_z B_z = |A|\,|B|$. It is necessary that the vectors shall be parallel to each other. So here the consistency of the relations demands a restriction on P_x, P_y, P_z, P_t, and it can easily be shown that this relation is

$$P_x w_x + P_y w_y + P_z w_z - c^2 P_t = 0.$$

Thus, when P_x, P_y, P_z, are assigned, a restriction is placed on P_t. In fact,

$$P_t = \frac{P_x w_x + P_y w_y + P_z w_z}{c^2}.$$

The quantity kw, or $w/(1-w^2/c^2)^{\frac{1}{2}}$ will be recognized as the relativity momentum of a particle of unit rest mass. However, to pursue further the details of relativity particle dynamics at this stage would take us into the realms of a detailed treatise on the theory of relativity, and would be out of place in a general survey of fundamentals. It will suffice to say that in seeking invariant relations to correspond to those of prerelativity physics, our procedure is usually to seek relations between 4-vectors (and sometimes between six vectors) which degenerate when the velocities are small to the old expressions familiar in prerelativity physics.

The Use of the Restricted Principle of Relativity. One is well justified in inquiring as to the ultimate use of the restricted theory of relativity. The answer is this: Suppose we have certain laws such, for example, as the Newtonian law of gravitation, which are very nearly correct, and we desire to formulate them in a more exact form, particularly to take account of what might happen in the case of velocities much greater than those we happen to have encountered in our initial formulation of the law. What have we to guide us? At first sight very little. We may place a very great restriction upon our possibilities by adjusting such modified laws as are proposed to be of a form invariant under relativity. Relativity cannot discover laws of physics for us, but it can act as a sort of supreme court to pass upon such laws as we propose; and, in this act it serves a useful function. Sometimes, the restriction which it imposes, when taken in conjunction with other required restrictions is so drastic as to leave only one final possibility open.

Then another use is illustrated by such a case as the following: Suppose a source fixed with respect to our axes is emitting monochromatic light of a certain frequency. By transferring to another system of axes moving with a velocity v in reference to the first set, it is possible to ascertain how the frequency of the light, wave-length, etc., are affected by a relative motion between the source and the observer. In fact, if we have the laws of any phenomenon given in a case in which one of the objects entering into the production of the phenomenon or indeed merely some identifiable point, has zero velocity, we can always, by relativity, find the corresponding law when the same object or identifiable point, is in motion. Of course, it must be noted that if bodies other than the one referred to are in motion, and if their velocities are involved in the statement of the law for the first case, relativity will have the power to pass upon that expression of the law as an invariant phenomenon or otherwise in that case also.

The Fundamental Content of the Restricted Theory: Removal of Certain Irrelevant Concepts. If we strip from the theory all the concepts incidental to its historical development, the fundamental outstanding dogma

---

Incidentally, we may remark that the subscript t in P_t is not intended to carry any special significance as relating P_t to the time. It is simply intended to denote the fourth component of the 4-vector, and is used because the fourth component of the 4-vector x, y, z, ict, is the t component.

which remains is that the laws of nature—the differential equations—shall be invariant under the transformation (1)–(4). *Once we have written down some proposed laws*, the test of this conclusion is a matter for pen and paper, and not for experiment. There is no need to form the concept of v as the velocity of some hypothetical observer who exists apart from ourselves. We can even take our proposed law and our transformation (1)–(4) to a mathematician and ask him to test its invariance without telling him what v is, or what c is. In order to deceive him in case he has read about the theory of relativity we can change v to another letter a, so that he will have no suspicion of its being a velocity, and his test will be the same as if he had read all about the classical visionary observers with their clumsy clocks and scales. We need not even tell him that x, y, z refer to space, and t to time. The test is purely an algebraical one. All of the fictitious observers, with their scales and clocks have evaporated, and there remains nothing in the theory of relativity but this question of the algebraical form of the laws in terms of x, y, z and t. But you may now begin to be a little doubtful as to how x, y, z and t are to be measured in order that the theory shall be true. Let there be any one way of measuring x, y, z and t for which the laws, when expressed in terms of those measures do conform to relativity. Then it is obvious that if a demon should take it into his head to alter the things by which you measured x, y, z, t, in such a way that you got different values from before, the laws which you would build up would, in general, not obey relativity. Or conversely, the demon may work the other way. It may be that with your clocks and scales or whatever you use, the laws do not conform to relativity, and the demon may make such adjustment in those instruments while you are making the measurements that the laws when chosen to agree with the numbers you finally write in your notebook do conform to the transformation. When we contemplate the rather arbitrary methods we do adopt for measuring space and time, we are lead to wonder whether it would not be well to state that the content of the theory of relativity is to the effect that there exists *some* way of assigning the numbers x, y, z, t in experiments such that the laws when adjusted to be consistent with measurements taken in that way shall be of a form invariant under the transformation (1)–(4). If we do this, any numbers which we might obtain and which did not lead to consistency with invariant laws are to be regarded as calling for some kind of correction, just as inequalities of temperature of the scales and clocks call for correction. Once we get one set of "corrected" measures, however, there will be an infinity of other measures obtainable by all possible transformations of the type (1)–(4), for which the laws will also be true. Now, I do not wish to raise here the question of whether we could always "correct" the measures in such a manner that in terms of the corrected measures, any law we might like to write down would become true, I wish simply to state that insofar as its working aspects are concerned the theory of relativity is only of value provided that it can successfully demand invariance of laws when expressed in terms of the measurements as we make them. On the other hand, starting with laws expressed initially in some definite set of coordinates, it is to be distinctly

understood that mathematical invariance of those laws under the Lorentzian transformation by no means implies that the coordinates for the transformed system would have any relation to actual measures of an observer travelling with a velocity v associated with transformation to that system. It is true that the principle of restricted relativity owed its formulation to a belief that the coordinates associated with the various systems corresponded to the actual measures; but, once formulated, the working content of the theory, involving as it does the mathematical invariance of the laws, is independent of this hypothesis.

It is usually assumed that if the Michelson-Morley experiment gives a positive result the theory of restricted relativity must be abandoned. I do not think that this conclusion can be substantiated. Let us imagine first a system of measurements such that physical laws when expressed in terms of them are mathematically invariant. Let us suppose these measurements to be made in a system, which, in the language of prerelativity days, would be regarded as fixed in space. Suppose that our earth is moving in relation to this "fixed" system with a velocity of about 18 miles per second, as the astronomers of old used to think to be the case, and let us suppose that some demon prevents those alterations of clocks and scales (alterations in relation to those of the fixed system) which the historical aspects of the theory of relativity would demand as a result of the velocity. Then all of our measurements will be a little different from what they would otherwise be; and, if we should test our laws by using the "erroneous" measure we should find them wrong. However, these errors will, in general, be beyond the limits of accuracy of our measurements. For since accurate experiments designed to test any such theory as the electromagnetic theory, for example, depend ultimately, in the last analysis, upon measurements of the distances between two points at a fixed time, or upon the difference between two times at a fixed point,[5] and since it is easy to show that such measurements will only be in error as the result of the supposed operation of the demon to the order v^2/c^2, it will be necessary to measure to this accuracy if the theories in question are to be found wrong. Thus such a great scheme of mathematical regularity as the electromagnetic theory, for example, lying as it were, with one part in 10^8 of our measurements will be discovered by us, although strictly speaking, we ought not to discover it. Then, we shall find the corresponding equations invariant under the transformation (1)–(4). This is not a matter of experiment, but one of pen and paper. Then we shall probably formulate the theory of relativity. We shall set out to test the way in which the mass of an electron varies with the velocity. We ought to find it different from that predicted by relativity, but we should have to measure the one part in 10^8 to discover the difference. Only when we do some experiments such as the Michelson and Morley experiment, which is capable of detecting one part in 10^8 shall we find the discrepancy. In the face of such a discrepancy, however,

[5] This is not true of "ideal," but impracticable experiments. The whole situation is discussed in detail in the writer's paper "Relativity and Aether Drift" in Phys. Rev. **35**, 336–346 (1930).

the valuable part of the theory of relativity still remains, the part which asserts that *there is a system* of measurement such that the laws when expressed in terms of it are invariant under the transformation (1)–(4); and, if there is one system, there is of course an infinity of others obtainable by all possible transformations of the type (1)–(4). The fact that our particular system of measurement may be as slightly different from this, as it would be if, in the languages of 30 years ago, the clocks and scales were not altered by motion, is a circumstance of little practical interest. The practical working feature of the principle of relativity, the feature which has to do with the forms of the laws in a *properly chosen* system of measurement, might still remain.

But it is natural to inquire, why should the laws of nature be invariant under this particular type of transformation? The answer to this question is to be sought in the general theory of relativity where the special theory finds a much more logical birthplace than in that old region populated by imaginary observers travelling with different speeds and measuring the velocity of light.

The General Theory of Relativity. It is not possible in an article of this length to attempt even an elementary account of the general theory in all its essential features, but I wish to refer to the subject for the purpose of emphasizing one or two features which are perhaps not always fully grasped. For one who is interested in relativity primarily in relation to electrodynamics, it may be advisable to omit the reading of this section.

The restricted theory of relativity made its appearance in 1905. In 1915, Einstein brought out his "General Theory of Relativity." This caused so much disturbance in the scientific world that the noise thereof got out into the popular world, which began therefore to hear for the first time of the *restricted* theory of relativity. Now, since the restricted theory of relativity has associated with it some of those more appealing paradoxes which delight the lay mind, and since the "man in the street" was quite unable to comprehend the general theory, we found a state of affairs in which the layman was talking about things more characteristic of the restricted theory under the impression that he was talking about the general theory.

To illustrate the matters of which we must now speak, let us consider a scaffolding extending up into the sky, with three sets of rods, mutually perpendicular to each other, and suppose that there are on these rods marks which are equally spaced, in the sense which you understand the phrase, and are labelled, 1, 2, 3, 4, etc. Suppose further that I am provided with a clock on whose face there are also marks which again are numbered 1, 2, 3, etc., giving me what I would call equal intervals of time. Suppose a sun and a planet move about in this scaffolding, and suppose I consider the Newtonian law as expressed by equations

$$m_a f_a = \frac{Gm_a m_b (x_b - x_a)}{\left[(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2\right]^{3/2}}$$

$$m_a g_a = \frac{G m_a m_b (y_b - y_a)}{[(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2]^{3/2}}$$

$$m_a h_a = \frac{G m_a m_b (z_b - z_a)}{[(x_b - x_a)^2 + (y_b - y_a)^2 + (z_b - z_a)^2]^{3/2}}$$

as a candidate for the expression of the facts of the case. What this means is, of course, perfectly obvious. It means that I take a very large number of readings of the position of the planet a and the sun b, for example, and the corresponding times, so that I am able to express the rate at which the position of the planet changes (its velocity), and the rate at which the velocity changes (the acceleration), and that I proceed to test whether the component accelerations are related to the corresponding space measures by this formula. I am not particularly concerned with discussing what correction I make on the time to allow for the time taken for the light from the planets to reach me. I simply suppose that I have decided upon some way of reading my clock, and possibly correcting its reading, for the purpose of obtaining a number t_a which I can associate with each value of x_a, y_a, z_a, which I measure for a, and a number t_b which I can associate with each value of x_b, y_b, z_b; and having written my data down on paper I proceed to test the law represented by the above equations. Suppose that I find that the law holds. Then that may well be the end of the matter. But, it will be a most astonishing thing if this does happen *exactly* for the reason which I shall try to make clear.

Suppose that I make each rod of my framework of scaffolding double, and that, leaving one frame which I shall call A as it is, I solder the points of intersection of the other, which I shall call B, and then distort the frame in an arbitrary way, leaving the same numbers attached to the points of intersection of the rods. Suppose further that I do a similar thing as regards the clock. In other words, suppose that I duplicate my original clock and leaving the old one as belonging to the system A, I modify the other one, which shall belong to B, in such a manner that it goes highly irregularly, and suppose that I modify my way of using its readings to assign times to the instants when objects coincide with certain intersections on the frames. Suppose I do this in a perfectly arbitary manner so that all that is left of the connection between the new time readings and the old is the fact that for every time reading that I assigned with my A clock as corresponding to the coincidence of a body with a point of intersection of my A frame, there corresponds one, and only one, time as recorded by my B clock for that same coincidence. I shall even go farther and imagine, that as time (recorded by the A clock if you wish) increases, the twisting of the B frame and the modification of the A clock alter. Moreover, I shall suppose that the B clocks and the method of calculation used in conjunction with them are continually changing, so that even the *relation* between the time readings recorded by the B and A clocks for coincidences of bodies with a given point of either frame changes with the time. Suppose I now use the numbers on the B frame to test the law given by the three above equations, which we supposed correct when the A

system was used. It is hardly to be expected that I shall find the law to hold now. But you will say, "of course not, the numbers are all wrong." Let us inquire a little further into this accusation against the numbers, however.

Suppose that when I asked you to imagine that the framework B became distorted I had asked you yourselves to take part in that distortion in every detail. Suppose, in fact, everything took part in this distortion. Then you would be totally unconscious of the change. For, every intersection of two or more lines would be preserved in that distortion in the sense that, if there were a certain identifiable molecule at the point of intersection in the old frame, that same molecule would be found at the corresponding point of intersection in the distorted frame. There might be a scale on which the numbers 1 and 2 and 3 were not separated by anything like equal distances as I viewed them, I who had not taken part in the distortion. As you viewed them, however, they would appear quite satisfactory; for your retina would be distorted, too, and the images of the number 1, 2, and 3 would fall on the same retinal cells as they fell on before the distortion, so that you will be quite satisfied with the picture. It is true that the A rods would appear crooked to you and the numbering on them would appear wrong. You might think that you could discover the inequality of distances between the marks on the B rods referred to above by using a movable scale which just fitted between the marks 1 and 2, and which you might suppose would fail to fit the space between 2 and 3. This supposition would be wrong, however, for, in terms of my standards of measurement, the rod would change in length as you moved it, and in such a way as to preserve the illusion. You would appear to me curiously deformed; but you could not test the crookedness of your backs because the straight edges you used would be crooked, too. You would not even look crooked to each other, because the distortion of the rays of light, and of the retinas of your eyes would be such as to bring the images of the various parts of your neighbors on to the same retinal cells as they fell on before the transformation. You might change your form as you moved about, but you would not know it for similar reasons. The same sort of thing would happen as regards time. You could not discover the irregularity in your clocks because any other clock you used to test it would suffer a like irregularity. You could not time it by your pulse beat because that would go irregularly, too. In fact, to you, everything associated with that crooked measuring system would appear correct. The most humiliating thing from my standpoint is that to you I should appear deformed and possessed of a highly irregular pulse. I should appear to alter my dimensions as I moved about; and, in fact, there would be absolutely nothing which I could do to prove that my system was undeformed and yours was deformed, that you could not do to prove that yours was undeformed and mine deformed. This being so, how do you know what the situation may be at this moment? If I miss out the introduction to this discussion, the picture of the deformation of the framework, etc., and simply start the story where we find ourselves with the difference of opinion as to which scheme of measures is right and which is wrong, you will readily see that there is no way of settling the

question. In fact, there is no question to be settled. There is no meaning to one being right and the other wrong. If you think there is, state it, and I will show how the advocate of either system will be able to prove in terms of your criterion that his is the right system. In fact, both systems are on a par with each other; and, since, as you will recall, the difference which I supposed to exist between the labelling of the two systems and the differences between the clocks was perfectly arbitrary, for it arose from an arbitrary distortion of the framework and arbitrary changes in the true measurements, all systems must be regarded as on an equal footing. Now if you can only eliminate from your consciousness the idea that one of these systems is right and all the others wrong, and you certainly ought to be willing to do this unless you can say what you mean by one being right, you will be in a position to admit that if we write down any law of such a kind that it is true when one measuring system (one method of attaching numbers to the events) is adopted and is untrue when another system is adopted, that law must have in it some properties of this system of measurement we have chosen. You will reply, "Of course it has. That is the object of physics, to choose a measuring system and then find laws expressed in terms of it." But if, as I have tried to emphasize to be the case, the system of measurement which we have chosen is a perfectly arbitrary one, should we expect to be able to express in terms of it, and in a simple form, any intrinsic regularity which there may be in Nature? Of course, we certainly can write down laws which are very approximately correct under certain conditions; but have we any right to expect simplicity in the particular set of measures we have chosen if the laws when transformed to any other set of measures, which are on equal footing with ours, assume another form; and, in general, a form which is almost infinitely complex?

It has become customary for the relativist to adopt the standpoint that laws which are not invariant in the sense of the general theory must have in them certain features which depend simply on the system of coordinates chosen; are therefore arbitrary, and can have no place in nature. It has been customary to maintain the stand that the exact laws of nature must conform, at least, to the criterion that they are invariant under a general transformation of coordinates.

It has been truly remarked that, in all physical measurements, the only things which are ultimately observed are coincidences of two or more marks. We say that the deflection of the spot of light was 3 cm; but, what was actually observed was that a certain dark line on the spot coincided with another dark line on the scale. Suppose that, taking all the numbers, x_1, y_1, z_1, t_1, x_2, y_2, z_2, t_2, . . . etc., which we have allotted to the various points in our experiments, including the numbers which have been allotted to the scales themselves it is obvious that if we should renumber all the points (both in space and time) in such a manner as to produce a one to one correspondence between the old system of numbering and the new, the coincidence between the old system of numbering and the new, the coincidence of two events (i.e. position and time) would be symbolized just as well by the

equality of the four numbers on the new system as by the four numbers in the old system. Thus, if we intend to imply in our laws no more than is contained in our observations there is no more reason for using one set of numbers than the other.

It is perhaps worthwhile pointing out that since our actual measures *do* make use of properties of the universe in their realization, the numbers which we assign to them do not have quite that arbitrary character which has been implied in the foregoing. The twisting of our scales, and the alterations of our clocks, etc., might well be criticized on the basis that a logical pursuance of the procedure would demand that the phenomena we were examining—the astronomical motions in the case cited—should also be made to suffer the same distortion, as otherwise we might destroy, unfairly, a sort of sympathetic understanding between the motions of the heavenly bodies and the ways of the scales and clocks of our actual experience. Thus, for example, it is a fact that the moon's disk is related to our rods in the sense that it has a simple shape—a circle—when measured with the rod. Again, the planetary orbits are nearly circles as measured by the rod, and so on.

We shall not attempt any further justification or criticism of the principle of invariance but shall now proceed to discuss its meaning in the form in which we meet it in the development of the subject.

The Significance of General Invariance. As a basis to which to pin our remarks, we shall take the general relativity law of planetary motion. For the immediate purpose in view it is unnecessary to enter into the question of how the law was reached.[6] It will suffice simply to take the law as it stands and discuss its significance. The law is as follows: A planet moves according to the differential equations,

$$\frac{d^2x_\alpha}{ds^2} + \frac{1}{2}\sum_{\mu=1}^{4}\sum_{\nu=1}^{4}\sum_{\lambda=1}^{4} g^{\alpha\lambda}\left(\frac{\partial g_{\mu\lambda}}{\partial x_\nu} + \frac{\partial g_{\nu\lambda}}{\partial x_\mu} - \frac{\partial g_{\mu\nu}}{\partial x_\lambda}\right)\frac{dx_\mu}{ds}\frac{dx_\nu}{ds} = 0 \qquad (54)$$

where

(1) x_1, x_2, x_3, x_4, correspond to the ordinary x, y, z, ict for the planet, so that (54) represents 4 parametric equations for x, y, z, ict, in terms of the arbitrary parameter s, which of course becomes eliminated in the process of getting x, y, z, as functions of t after the equations are solved.

[6] Our avowed purpose is to avoid encumbering our comments on the significance of the general theory with the question of how the invariant laws are arrived at. For the purposes of the discussion it is sufficient to accept (54) and (55) as having fallen from heaven. However, for the benefit of those unfamiliar with the subject, it may be well to add the following: (54) is the equation of a Geodesic. It is the analytical equivalent of stating that the particles move in such a way that

$$\delta\int ds = 0 \qquad (54.1)$$

where ds is given by (56). The invariance of (54) is guaranteed by the fact that $\delta\int ds = 0$ is the condition for a maximum or minimum of the integral, and (54.1) is an invariant relation. Even if the particle moved in one system of coordinates in a straight line with constant velocity, *i.e.*, if it moved so that its spacial coordinates changed linearly with the time, we could disguise this simple motion into all sorts of complicated motions by expressing it in terms of other

(2) The sixteen[7] $g_{\mu\nu}$'s are functions of x_1, x_2, x_3, x_4, concerning which, all we say is that they are solutions of a certain definite set of 16 partial differential equations in x_1, x_2, x_3, x_4, which equations could be written down here in full if there were space, but which are symbolized by writing

$$G_{\mu\nu} = 0. \tag{55}$$

(3) The quantity $g^{\alpha\lambda}$ is the minor of the term $g_{\alpha\lambda}$ in the determinant formed from the sixteen g's.

To the foregoing we must add a very important remark. If we make an arbitrary point to point transformation of coordinates,

$$\begin{aligned} x_1 &= f_1(x_1', x_2', x_3', x_4') \\ x_2 &= f_2(x_1', x_2', x_3', x_4') \\ x_3 &= f_3(x_1', x_2', y_3', x_4') \\ x_4 &= f_4(x_1', x_2', x_3', x_4') \end{aligned}$$

(54) and (55) come back to exactly the same form in dashed coordinates and in dashed g's with this important understanding—that the *dashed g's are related to the undashed g's by the transformation corresponding to that of a covariant tensor of the second rank.* Now in case the reader is unfamiliar with this language, it will form the analytical equivalent of it to say that the dashed g's may be obtained from the undashed g's in the following manner: Construct the quantity ds^2, defined as

$$ds^2 \equiv \sum_{\mu}\sum_{\nu} g_{\mu\nu} dx_\mu dx_\nu \tag{56}$$

where the g's are of course functions of dx_1, dx_2, dx_3, dx_4. Now make the transformation. It is easy to see that ds^2 in the new coordinates assumes the form of a homogeneous quadratic function in $dx_\mu' dx_\nu'$; and the coefficient of $dx_\mu' dx_\nu'$ is the quantity $g_{\mu\nu}'$ referred to above.

---

coordinates related to the first set by some general transformation. We might even be lead to wonder whether the complicated motions of the heavens could not all be represented as straight lines with constant velocity in some suitably chosen system of coordinates. If such were the case, the law of astronomy would be that a particle moved according to (54) where the g's were the sort of g's which could have come from the g's 1,1,1,1, by some transformation or other. The condition that these sorts of g's must satisfy can be expressed as a set of differential equations governing the g's, which differential equations are usually written

$$B^{\tau}_{\mu\nu\sigma} = 0.$$

The notation need not detain us. It will suffice to say that it can be shown that in all the possibilities available under such a situation, there are none which will give a motion approximating to the planetary motions. The next question is whether there are any other differential equations which the g's could satisfy, differential equations which would deny that they could have come from the g's 1, 1, 1, 1, by transformation, but which would permit solutions which when substituted in (54) made it fit the facts. The answer is that there is such a set, and (55) is the set in question.

[7] Owing to the fact that in practical cases $g_{\mu\nu} = g_{\nu\mu}$ the number of the g's reduces to 10. However, we do not wish to encumber the exposition with this fact at the present time.

In other words, the relation between the dashed and the undashed g's is that implied in

$$\sum_{\mu\nu} g_{\mu\nu} dx_\mu dx_\nu = \sum_{\mu\nu} g_{\mu\nu}' dx_\mu' dx_\nu'. \tag{57}$$

Now let us elaborate a little on the foregoing. In ordinary dynamics, we have equations of motion analogous in a sense to (54), but of the form, for example

$$\frac{d^2 x_\alpha}{ds^2} - \text{Force} = 0. \tag{58}$$

Our business is to solve the four equations, assign the initial values of the x_α and the dx_α/dt, and the problem is finished. Equation (54) is not as definite as (58), however. Before we can solve it we have to assign not only initial values of the x_α and dx_α/dt, but we have to assign the g's which occur in the parenthesis. For these we look to equation (55), and all this tells us is that we can have any g's we like, so long as they satisfy (55). In fact, these two equations (54) and (55) together are much more general—they say much less than an equation like (58). That is the penalty we have to pay on their behalf for their luxurious heritage of invariance. Of course, when we start to solve (55) we again need to *assign* certain values of the g's, certain boundary values as they are called. In practise we do not usually do this. What we do is to get *some* set of g's which satisfy (55) and which when substituted in (54) make that equation give final results approximately like the Newtonian equation symbolized by (58). When we have done this we can state that our ultimate solution of the problem has been built up as a special case of a set of equations having the characteristic of invariance. Indeed, there would be nothing *surprising* if we had found that we could have reproduced exactly our old Newtonian equation as such a special case. The result would be very instructive to our thinking if we could. For in that case, one person could look at our final equations and say they are not invariant at all. They are not invariant even for the simple transformation of the restricted theory let alone the general transformation of the general theory. And we should say to him, "You are looking at invariance in the wrong way. You must not boil the parenthesis of (54) down to an exact expression in the coordinates and then ask for invariance. The equations are not as invariant as all that. The parenthesis will not appear in the same form in the coordinates on transformation. It will appear in the same form in the derivatives of the g's with the understanding that the dashed g's are related to the undashed g's according to the principle stated above, and implied in (57)."

Let us view the matter from another angle. Suppose I should consider two beings A and B who start to study an identical physical problem—an actual planetary motion in two different sets of coordinates.

A will, or should, assign his boundary values to the g's and solve (55). For the moment I am not interested in saying how he is to assign the boundary values. Next he will substitute for the g's in (54). He will then assign the necessary initial (or boundary) values for the appropriate quant-

ities concerned in (54) and proceed to his solution. Now B will start with equations like (54) and (55) with everything dashed; but, in order that he shall get the same solution as A[8] it will be necessary that first he shall assign not the same but corresponding boundary values in (54), corresponding in the sense that one is obtained from the other by the transformation which takes us from A's coordinates to B's. This is not all, however. For, before B gets his equation (54) he must solve (55) and here he must assign the boundary values to the dashed g's so that the boundary values of the dashed g's are related to the boundary values of the undashed g's assigned by A, by the law implied in (57). But how is B to know what boundary values A assigned? There is nothing simple which he can measure to get his boundary g's as he could measure the position and velocity of the planet to assign boundary values to (54). The answer to the conundrum is that really neither A nor B have any immediate way of assigning boundary values to the g's. They must leave them arbitrary—at least, if they are going to be respectable mathematicians and avoid the guessing process referred to above. They must solve (55) with arbitrary boundary g's. They must let these arbitrary quantities get into (54). They must assign the boundary values peculiar to (54) by direct observation, and solve (54). The solutions will then still be indefinite to the extent of the arbitrariness of the boundary values of the g's still remaining in that solution. The whole content of the truth of the theory now lies in the possibility that A shall be able to assign such boundary values as will make the solution of (54) fit the facts. He will have to take a lot of points on the orbit, determine the boundary values in question and then see if his solution fits all the other points on the orbit. It is obvious that in general there will only be one set of boundary values (if any) which will do this. Moreover, we may feel quite certain that if B does the same thing, the boundary values of the dashed g's assigned by B will be related to those assigned by A by the relation implied in (57), for that must follow as a mathematical necessity.

Perhaps an apology is needed for this long digression on the significance of invariance which is usually dismissed in so few words. I tender such an apology to those whose conscience permits them to accept it.

The Relation between the Restricted Theory and the General Theory. One very important point remains. We have interested ourselves in the quantity

$$ds^2 \equiv \sum_{\mu}\sum_{\nu} g_{\mu\nu} dx_\mu dx_\nu \tag{59}$$

Now, if we consider the quantity

$$dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2 \tag{60}$$

and make an arbitary transformation of coordinates it assumes the form (59) as is very well known and easily seen. However, even with an *arbitrary* transformation, we cannot produce from (60) all possible expressions like (59).

[8] The same in the sense that the motion calculated by B will be exactly the same as that obtained by superposing the transformation of coordinates on A's calculation.

We cannot make the g's any function of x_1, x_2, x_3, x_4 we like by an arbitrary transformation. Now there is a set of coordinates (we call it an Euclidean set) in which far away from what we call matter a material particle moves in a straight line with constant velocity. In other words this particle obeys the law

$$\frac{d^2x_\alpha}{ds^2} = 0$$

which is a particular case of (54), when the parenthesis is zero, as it would be if the g's were constants. It is consequently customary to limit our interest to those g's which we get as solutions of (55) and which at infinite distance[9] from matter assume such functions of the coordinates that it is possible to find a transformation which will transform them to the g's, 1, 1, 1, 1, characteristic of (60). How if we are in a set of coordinates in which the g's are 1, 1, 1, 1, i.e., in which

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 + dx_4^2$$

we know that a *Lorentzian transformation of the restricted theory of relativity*, i.e., the transformation (1)–(4) will leave those g's equal to 1, 1, 1, 1. This is the fundamental characteristic of the Lorentzian transformation. One of the best known ways of determining the transformation is to determine it as the transformation which leaves $dx^2+dy^2+dz^2-c^2dt^2$ invariant. Hence for this transformation, the g's remain 1, 1, 1, 1, i.e. they certainly remain constants. Hence if there be one set of coordinates in which the g's disappear from the equation owing to their being constants whose derivatives consequently are zero, then they will disappear in any other set of coordinates obtained from the first by a Lorentzian transformation. Hence, if there be any set of equations (containing derivatives of the g's) which are invariant in the sense of the general theory of relativity, then the equations obtained by putting the derivatives of the g's equal to zero will be invariant in the more drastic sense of the restricted theory of relativity. It would thus seem that the restricted theory of relativity finds a more logical origin as a special case of the general theory, an origin which emphasizes its bearing on the mathematical forms of the expressions, than it finds in arguments having to do with different observers moving relatively to each other and performing certain hypothetical experiments.

The Significance of the Velocity of Light in the General Theory. The customary method of approach to the general theory, is a sort of generalization of the restricted theory. The velocity of light c, which, again in the customary approach, makes its appearance through the activities of hypothetical observers crawls into the general theory by a sort of back door, which leaves one with the sort of feeling of surprise in the end that it should

[9] The significance of infinity is ambiguous in an arbitrary unknown set of coordinates. All that we ought to say is that it is possible to assign a set of coordinates such that the g's when expressed in terms of them assume forms transformable to 1, 1, 1, 1 at infinite values of those coordinates.

be found in the general theory at all. By removing the artificial and unnecessary steps peculiar to the historical development, it is possible to place these matters on a more satisfactory basis. The problem of particle motion will provide a sufficiently general basis to illustrate the idea.

Starting with no thoughts of light in our minds, let us proceed from (54) and (55). As is well known, and as may be verified from all the classical treatises on the subject, one solution of (55) for the g's in terms of r, θ, ϕ, makes the ds^2 of (56) take the form

$$ds^2 = -\gamma^{-1}dr^2 - r^2d\theta^2 - r^2\sin^2\theta d\phi^2 + c^2\gamma dt^2 \tag{61}$$

where $\gamma \equiv 1-2m/r$, m, and c being merely constants. Formally, there is no reason to recognize r, θ, ϕ as polar coordinates, although the particular solutions of (55) which have been chosen have been chosen with the idea that ds^2 comes out of a form closely resembling that which we have been accustomed to speak of as the line element in the polar coordinates. I do not wish to pursue this matter because it would lead us into fields interesting in themselves, but calculated to obscure the points really at issue. Since any transformation of coordinates will give from the g's of (61) another set of g's which are also solutions of (55), it is perfectly obvious that we could omit the c^2 in (61), and indeed could modify the expression by writing $r=ar'$, $\theta=b\theta'$, $\phi=k\phi'$ if we wished, and still have g's which were solutions of (55). But in order that we may have an ultimate solution for the planetary motion it is necessary that we shall not take *any* solutions of (55), but the *right* ones. In including c^2 in (61) we have anticipated to some extent the needs of the final solution. It is now a question of mere algebra, a question which need not detain us, and which will be found elaborated in any of the standard books, to show that with the g's specified in (61), equation (54) may be shown to result in an orbit which is in a plane of constant θ (taken arbitrarily as $\pi/2$), and in which the motion is governed by

$$\frac{d^2u}{d\phi^2} + u = \frac{m}{h^2} + 3mu^2 \tag{63}$$

$$r^2\frac{d\phi}{dt} = h\left[c^2\gamma - \gamma^{-1}\left(\frac{dr}{dt}\right)^2 - r^2\left(\frac{d\phi}{dt}\right)^2\right]^{1/2} \tag{64}$$

where $u \equiv 1/r$, and where h is a constant of integration.[10]

Integrating (63), with approximations for convenience (though not of necessity)[11] we arrive, in the usual way at

[10] The form usually given for the equation is $r^2d\phi/ds=h$. We have substituted for ds/dt from (61).

[11] Our wording is here a little careless, since there has been no mention of the sizes of quantities to warrant speaking of approximations. Our association of r, θ, ϕ, with polar coordinates is purely arbitrary, founded on a determination to find *some* solution of (54) and (55) which shall ultimately give an orbit approximating that which we find in terms of those coordinate measures which we actually make and which we call polar coordinate measures. Having made our association of coordinates, the ultimate association of the constants m and h with measured

$$u = \frac{m}{h^2}[1 + e\cos(\phi - \omega - \delta\omega)] \tag{65}$$

where e and ω are constants for the orbit, but

$$\delta\omega = \frac{3m^2}{h^2}\omega. \tag{66}$$

Apart from the small term $\delta\omega$, an examination of the orbit would serve to determine only m/h^2, and not m and h separately. The term $\delta\omega$ serves, however, to fix m^2/h^2, so that by its inclusion, both m and h are determined from experimental observations. The determination of h now enables c to be determined by the substitution of measured values of $d\phi/dt$ in (64).[12] The type of variation of ϕ with the time necessitated by (64) is of course, a requirement of the theory. In practise, since c turns out to be so large compared with dr/dt or $rd\theta/dt$, and since γ is very nearly equal to unity, (64) may be written with sufficient approximation as

$$r^2 d\phi/dt = hc$$

Without knowledge of h it would have been impossible to determine c. This quantity c thus finds its natural origin in the departure of the motion from the Newtonian law, and it plays no part in the principal part (the Newtonian part) of the motion.

But now there is something else. It can be shown that if, in the sense defined by

$$v = \left[\left(\frac{dr}{dt}\right)^2 + r^2\left(\frac{d\theta}{dt}\right)^2 + r^2\sin^2\theta\left(\frac{d\phi}{dt}\right)^2\right]^{1/2} \tag{67}$$

a particle has a velocity v which is less than c, equations (63) and (64) will never allow it to attain a velocity greater than c. The existence of a maximum value c for the velocity attainable by a particle reminds us of the statement that no particles can move faster than the velocity of light, so we are tempted to inquire whether the velocity of light (understood as measured far from matter, of course,) will serve for the quantity c, which had its origin in the deviations from the Newtonian law of planetary motion. We find that the velocity of light will so serve.

Again, since the maximum velocity which a particle can attain is equal to the velocity of light, we are tempted to try whether a ray of light will follow the path of a particle which (at infinity of course) travels with the velocity of light. In this way we are led to the idea of the bending of light by the sun in a manner which may appear more suggestive to some than

quantities for this orbit insures the validity of neglecting $3mu^2$ compared with m/h^2 as a first approximation, and using the approximate solution thus obtained in the term $3mu^2$ itself. It is in the sense of anticipating this that our phrase concerning approximation has correct significance.

[12] dr/dt may be expressed, of course, in terms of $d\phi/dt$ through the equation of the orbit.

one which leads to it through what experts in the subject will recall as the Principle of Equivalence.

The Symmetry of the Line Element at Infinity. The fact that when $r = \infty$, (61) reverts to a form which when expressed in cartesian coordinates becomes

$$ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2 \tag{68}$$

may, at first sight, seem to carry with it the conclusion that the restricted theory must necessarily hold at infinity (at least as far as the laws under discussion are concerned). In accordance with the principles enunciated in the section on "The Relation between the Restricted Theory and the General Theory," this conclusion would indeed be expected if we were certain that the coordinates utilized in (68) represented the actual measures of observers. The fact that (68) has resulted from putting $r = \infty$ in (61) does not necessarily warrant this conclusion as to the coordinates. For suppose that, having written down (61), we had, before proceeding made a transformation such as[13]

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t. \tag{69}$$

The new g's appropriate to the new expressions of the line element would be solutions of (55), for *no* transformation will rob them of this property. In proceeding to our solution in these coordinates, we should of course obtain exactly the solution usually obtained, i.e. the solution corresponding to (63) and (64), but transformed to the dashed coordinates specified in (69). We should obtain exactly the result which would be expected by a physicist of mid-Victorian times who had somehow or other been lead empirically to the equations of motion (63) and (64) as applicable to the sun at rest, and had then supposed the sun to move through space with a velocity v. When we proceed to infinity with our x', y', z', t', coordinates, however, we no longer obtain a line element of the form (68); and, all that the *general* theory predicts as regards a *restricted* theory is that if we make the transformation back from x', y', z', t', to x, y, z, t, we shall find ourselves in a set of coordinates such that the laws expressed in terms of them will be invariant under the Lorentzian transformation. They will not be so invariant to one who has expressed them in terms of x', y', z', t', however.

Electrodynamics

Statement of the Fundamental Equations. The laws of electrodynamics are embodied in the well-known scheme of equations[14]

$$\frac{1}{c}\left(\rho u + \frac{\partial E}{\partial t}\right) = \text{curl } H, \tag{70}$$

$$\rho = \text{div } E, \tag{71}$$

[13] We should really express this transformation in terms of polar coordinates, but for simplicity have left it in cartesians. This fact will cause the reader no difficulty.

[14] The units here used are the Heavisidian units, in which the unit of charge and unit magnetic pole are respectively $1/\sqrt{4\pi}$ times the electrostatic and magnetostatic units.

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t} = \text{curl } \boldsymbol{E}, \tag{72}$$

$$0 = \text{div } \boldsymbol{H}, \tag{73}$$

together with a fifth relation whose aim is to describe the law of motion of an electron. This law states that the electron is a spherical shell of charge when viewed in a system of axes in which its center has zero velocity, and that in this system of axes, its acceleration and the other time derivatives of its motion are determined in such a manner that

$$\iiint \rho \boldsymbol{E} d\tau = 0. \tag{74}$$

In the above equations, ρ is the charge density, $\boldsymbol{E}$ and $\boldsymbol{H}$ the magnetic fields $\boldsymbol{u}$ the velocity, $d\tau$ an element of volume, and c a constant, which experiment shows to be equal to the velocity of light in vacuo.

It is customary to express (74) in another form. We first split the integral into two parts, one pertaining to the field $\boldsymbol{E}_i\boldsymbol{H}_i$, produced by the charge on the electron itself, disentangled from the total field in a manner to be explained later and the other, the field $\boldsymbol{E}_0\boldsymbol{H}_0$, due to all the remaining charges of the system. Equation (74) may be then written

$$\iiint \rho \boldsymbol{E}_0 d\tau = -\iiint \rho \boldsymbol{E}_i d\tau. \tag{75}$$

It is possible to calculate $\boldsymbol{E}_i$ in terms of the acceleration $\ddot{\boldsymbol{S}}$, rate of change of acceleration, $\dddot{\boldsymbol{S}}$, and higher time derivatives of the motion, so that expressing the right hand side of (75) in this way, the equation assumes the form

$$\iiint \rho \boldsymbol{E}_0 d\tau = \frac{e^2}{6\pi a c^2}\ddot{\boldsymbol{S}} - \frac{e^2}{6\pi c^3}\dddot{\boldsymbol{S}} + \cdots, \text{ etc.} \tag{76}$$

where a is the radius of the spherical shell, and e the electronic charge.

The left hand side of (75) being called the external force on the electron, the quantity $e^2/6\pi a c^2$, which is the coefficient of the acceleration on the right hand side is appropriately known as the mass of the electron, it being understood that (76) is applicable only to the case when the velocity of the electron is zero in the system of axes to which $\boldsymbol{E}$ and $\boldsymbol{H}$ are referred. In many cases it is customary to neglect altogether the terms on the right hand side of (75) other than the first, and to assume moreover that the external field is constant over the electron. In this case, (76) assumes the form

$$e\boldsymbol{E}_0 = \frac{e^2\ddot{\boldsymbol{S}}}{6\pi a c^2}. \tag{77}$$

In extending the law of motion of the electron to cases where the velocity of the electron is not zero in the system of axes in which $\boldsymbol{E}$ is measured, it is customary to invoke the assumption of the restricted theory of relativity

which, at any rate for the case where $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ may be assumed constant over the electron, leads via (75) to the result[15]

$$e\left(\boldsymbol{E}_0 + \frac{[\boldsymbol{w}\boldsymbol{H}_0]}{c}\right) = \frac{d}{dt}\left\{ m_0 \boldsymbol{w} \Big/ \sqrt{1 - \frac{\boldsymbol{w}^2}{c^2}} \right\} \qquad (78)$$

$$+ \text{ terms involving higher time derivatives}$$

where $m_0 = e^2/6\pi ac^2$, and $\boldsymbol{w}$ is the velocity of the center of the electron. For the case of an acceleration parallel to the velocity, the first term on the right hand side of (78) assumes the form $m_0\dot{w}/(1-w^2/c^2)^{3/2}$, and for the case of an acceleration perpendicular to the velocity it assumes the form $m_0\dot{w}/(1-w^2/c^2)^{\frac{1}{2}}$. We consequently speak of

$$\text{The Longitudinal Mass } \frac{e^2}{6\pi ac^2}\left(1 - \frac{w^2}{c^2}\right)^{-3/2}$$

and

$$\text{The Transverse Mass } \frac{e^2}{6\pi ac^2}\left(1 - \frac{w^2}{c^2}\right)^{-1/2}.$$

On the foregoing view, the quantity which is really measured when the mass is measured for small velocities is the quantity $e^2/6\pi ac^2$, and the value of a is then calculated as the radius of a sphere which would give rise to that mass.

From the standpoint of classical electrodynamics and apart from quantum phenomena, these equations are regarded as applying universally outside, and even inside matter, so long as proper account is taken of the fields and charges arising from the atomic structure. The quantities usually denoted by $\boldsymbol{B}$ and $\boldsymbol{D}$, as distinct from $\boldsymbol{H}$ and $\boldsymbol{E}$ arise in the process of building up equations limited to application on a macroscopic scale. From the standpoint of the electron theory such equations owe their existence to an endeavor to amplify the simple equations to enable them to take account, at any rate approximately, of the presence of matter as far as its macroscopic effects are concerned, without invoking a detailed knowledge of the intramolecular structure. Since the present article is concerned with electrodynamics, mainly in its fundamental aspects, we shall not give any further consideration to the macroscopic equations.

Our statement above concerning the universal applicability of the equations for free space was qualified by the phrase "Apart from quantum phenomena." This reservation is of course very vital. For, in the last analysis, all phenomena are manifestions of atomic phenomena even when disguised in macroscopic form. The motions in a motor are manifestations of what are ultimately atomic phenomena. The best that can be said for the universal applicability of the equations is that while we do not regard them as applic-

[15] If $\boldsymbol{E}_0$, $d\tau$, and $\boldsymbol{H}_0$ are not constant over the electron, the 'eft-hand side of (77) assumes the form $\iiint \rho(\boldsymbol{E}_0 + [u\boldsymbol{H}_0]/c)d\tau$, but with the various elements of the integral calculated at different times.

able to all cases of interactions between electrons, they do give, when averaged for such large scale phenomena as those concerned in the motor and other appliances of engineering practice, results which are correct and which are therefore the same as would be obtained by averaging from the fundamentally correct laws.

The status of the equations of electrodynamics in a scheme of modern quantum theory is hardly yet established in complete logical form. We shall, however, leave this matter for the present and discuss first the historical development and significance of the equations in relation to what we may call the classical period immediately prior to the advent of the quantum theory.

Historical Survey.—At the time when Maxwell commenced his epoch-making investigations the known laws of electrodynamics were summed up in conclusions which are the analytical equivalent of the following:

$$4\pi I = \oint \boldsymbol{H}_s ds, \tag{79}$$

expressing the fact that the line integral of the magnetic field around a closed circuit encircling a current I is equal to $4\pi I$.

Then we have the relation[16]

$$-\frac{d}{dt}\int\int \boldsymbol{H}_n dS = \oint \boldsymbol{E}_s ds, \tag{80}$$

expressing Faraday's law, that the rate of change of the surface integral of the normal component $\boldsymbol{H}_n$ of the magnetic field taken all over a cap bounded by a circuit is equal to the negative of the line integral of the complete electromotive intensity[17] taken all around the circuit. Finally we have a law having to do with the mechanical forces on a circuit carrying a current, and equivalent to the statement that in the case of a number of circuits expressed by generalized coordinates q_r, the generalized force F_r necessary to hold the circuit in position against the actions of the other circuits is[18]

$$F_r = -I_r \frac{\partial}{\partial q_r}\int\int \boldsymbol{H}_n dS\,. \tag{81}$$

[16] We are confining ourselves to the case of free aether, so that we use $\boldsymbol{H}$ instead of $\boldsymbol{B}$ in (80). The introduction of $\boldsymbol{B}$ and $\boldsymbol{D}$ into the discussion would carry with it more complexity of an irrelevant kind than is worth while for the purpose of the present historical survey. Our units are here, of course, the ordinary electromagnetic units, since the introduction of Heavisidean here would add needlessly to the apparent artificiality of the equations of the pre-Maxwell pe riod.

[17] Strictly speaking, the pre-Maxwellian era was acquainted only with the right hand side of (80) taken as a whole, i e. with the electromotive force. The dissection of this electromotive force into the line integral of a quantity E_s anticipates the later stage of development in which the thought is concentrated on the force on an element of the electric current.

[18] The directions are such that viewing the circuit from that side from which the current appears to flow clockwise, the positive direction of the normal is drawn from the observer through the cap.

The negative of this quantity, $f_r(=-F_r)$ is the generalized force on the circuit due to the other circuits, so that:

$$f_r = I_r \frac{\partial}{\partial q_r} \int\int \boldsymbol{H}_n dS, \tag{82}$$

which is the form in which the expression is usually thought of.

Equation (82) is of course the analytical equivalent of the law obtained by calculating the forces between the current circuits by treating them in terms of their equivalent shells, and it is also the analytical equivalent of the statement that the forces on a current circuit as a whole, due to the other circuits is the same as that obtained by attaching to each element of length ds of the conductor a force δf_r, given by

$$\delta f_r = [I_r \cdot \boldsymbol{H}]\delta s \tag{83}$$

Maxwell set himself the task of moulding equations (80) and (81) into a form which made them appear as the outcome of general dynamical laws.

In the hands of Maxwell, equation (79) amounted practically to a definition of $\boldsymbol{H}$ in terms of the currents. It is not difficult to show that (79) contains as an analytical consequence, that insofar as $\boldsymbol{H}$ is determined by the current alone, the flux of $\boldsymbol{H}$ through any circuit denoted by subscript r, is given by

$$\int\int \boldsymbol{H}_n dS = L_{1r}I_1 + L_{2r}I_2 + \cdots L_{rr}I_r + \text{etc.} \tag{84}$$

where the L's are the ordinary coefficients of self and mutual induction, and are perfectly definite functions of the geometry of the system, given by

$$L_{mn} = \int\int \frac{\cos\phi ds_m ds_n}{r} \tag{85}$$

ϕ being the angle between the elements ds_m and ds_n of the circuits concerned.

The Faraday law (80), for the circuit r becomes

$$-\frac{d}{dt}(L_{1r}I_1 + L_{2r}I_2 + \cdots L_{rr}I_r + \text{etc.}) = p_r \tag{86}$$

where p_r is the electromotive force $(=\int \boldsymbol{E}_s ds)$ produced in the circuit as a result of the varying currents in the circuit and the other circuits. Or, if $P_r = -p_r$ is the generalized external force necessary to result in the condition of changing currents in the circuit r.

$$\frac{d}{dt}(L_{1r}I_1 + L_{2r}I_2 + \cdots L_{rr}I_r + \text{etc.}) = P_r. \tag{87}$$

Again, the law (81) for the generalized force acting on the circuit r becomes[19]

[19] Of course, $\partial L_{rr}/\partial q_r$ is zero, and is merely included for symmetry.

$$F_r = -I_r\left(I_1\frac{\partial L_{1r}}{\partial q_r} + I_2\frac{\partial L_{2r}}{\partial q_r} + \cdots I_r\frac{\partial L_{rr}}{\partial q_r} + \text{etc.}\right). \tag{88}$$

Thus the problem is to represent (87) and (88) as a consequence of generalized dynamics. The procedure is now almost obvious. Lagrange's equations for generalized coordinates are of the form:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_m}\right) - \frac{\partial T}{\partial q_m} = \Phi_m \tag{89}$$

where Φ_m is the generalized force associated with the coordinates q_r. Suppose, in addition to the mechanical coordinates, we think of electrical coordinates which we shall make appear only in our kinetic energy function through their velocities, which we shall associate with the various currents I_1, I_2 . . . . . Suppose in fact, we take for T the expression:

$$T = \tfrac{1}{2}L_{11}I_2^2 + L_{12}I_1I_2 + L_{13}I_1I_3 + \cdots \tfrac{1}{2}L_{22}I_2^2 + L_{23}I_2I_3 + L_{24}I_2I_4 + \cdots \text{etc.} + T_m, \tag{90}$$

where T_m is the expression we should have for the kinetic energy of the matter of the wires in absence of any currents. Then, applying Lagrange's equations to the electrical coordinate represented as regards its velocity by I_r, we obtain (since T does not involve the electrical coordinates)

$$\frac{d}{dt}(L_{1r}I_1 + L_{2r}I_2 + \cdots) = P_r \tag{91}$$

which is the same as (87) the equivalent of the Faraday law. And, by applying the equations to the mechanical coordinate q_r, we obtain

$$-I_r\left(I_1\frac{\partial L_{1r}}{\partial q_r} + I\frac{\partial L_{2r}}{\partial q_r} + \cdots I_r\frac{\partial L_{rr}}{\partial q_r}\right) + \frac{d}{dt}\left(\frac{\partial T_m}{\partial \dot{q}_r}\right) - \frac{\partial T_m}{\partial q_r} = \Phi_r. \tag{92}$$

The part involving T_m is simply the part of Φ_r necessary to accelerate the brute matter of the conductors. In the case where the conductors are held fixed, Φ_r represents what we have called F_r and (92) becomes

$$-I_r\left(I_1\frac{\partial L_{1r}}{\partial q_r} + I_2\frac{\partial L_{2r}}{\partial q_r} + \cdots I_r\frac{\partial L_{rr}}{\partial q_r}\right) = F_r \tag{93}$$

which is the same as our law (88), which is the equivalent of (81).

We may regard the achievement of Maxwell as that of finding the necessary expression for T to permit of the moulding of the equations into the dynamical form in the sense we have described.

Maxwell next endeavored to extend the theory to open circuits, and to the motions of charges. Writing (79) in differential form it becomes

$$4\pi \boldsymbol{i} = \text{curl } \boldsymbol{H} \tag{94}$$

where $\boldsymbol{i}$ is now the current density, it being understood, however, that the stream lines of the flow continue themselves into complete circuits.

The thought is that $\boldsymbol{i}$ may be regarded as a flow of some density with a velocity $\boldsymbol{u}$, and that we may be able to write

$$\frac{4\pi\rho\boldsymbol{u}}{c} = \operatorname{curl} \boldsymbol{H}. \tag{95}$$

The factor c is included because, if ρ is in electrostatic units, our thought of the similarity between what has been called current density and $\rho\boldsymbol{u}$ has implied only a proportionality between $\rho\boldsymbol{u}$ and $\boldsymbol{i}$, since the unit of $\boldsymbol{i}$ in equation (94) has been chosen so that $\boldsymbol{H}$ shall represent the force on unit pole in such cases where that test is applicable. c is in fact, the ratio of the electrostatic to the electromagnetic unit of charge.

With equation (95) however, we immediately encounter a difficulty; for, since div. curl $\boldsymbol{H}=0$, that equation would imply

$$\operatorname{div} \rho\boldsymbol{u} = 0.$$

But, if we are to regard electricity in the light of an indestructible entity, its motion must be controlled by the equation of continuity

$$\operatorname{div} \rho\boldsymbol{u} + \frac{\partial\rho}{\partial t} = 0 \tag{96}$$

so that (95) must be regarded as impossible for even such a simple case as that of a moving charge of finite spacial extent, since here the density at any point changes with the time.

In order to escape this difficulty, Maxwell generalized (95), which is strictly true only when $\partial\rho/\partial t=0$, to the equation

$$\frac{4\pi}{c}(\rho\boldsymbol{u}) + \frac{1}{c}\frac{\partial\boldsymbol{E}}{\partial t} = \operatorname{curl} \boldsymbol{H} \tag{97}$$

which, on taking the divergence of both sides gives

$$4\pi \operatorname{div} \rho\boldsymbol{u} + \frac{\partial}{\partial t}\operatorname{div} \boldsymbol{E} = 0 \tag{98}$$

so that, provided that

$$4\pi\rho = \operatorname{div} \boldsymbol{E} \tag{99}$$

(98) becomes

$$\operatorname{div} (\rho\boldsymbol{u}) + \frac{\partial\rho}{\partial t} = 0$$

so that the situation is now consistent with the equation of continuity.[20]

The quantity $\boldsymbol{E}$ occurring in (99) is of course, measured in electrostatic units, while the $\boldsymbol{E}$ in (80) is associated with a current I measured in the same

[20] The generalization on (95) implied in the addition of the term involving $\partial E/\partial t$, is of course not the only one possible. We might add in addition any vector whose divergence was zero without violating (96).

units as are employed in (79). Hence, if $\boldsymbol{E}$ is to be in electrostatic units, we must divide the left-hand side of (80) by c; and, expressed in differential form, and applied to a fixed circuit of infinitesimal dimensions, the equation would then become

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t} = \text{curl } \boldsymbol{E}. \tag{100}$$

It will here be noted that we have a partial time derivative in place of the total time derivative of (80) which latter equation is applicable to a moving circuit as well as to a fixed one.

Equation (82), with the current (there in electromagnetic units) regarded in the light of a generalization to moving charge density becomes the equivalent of stating that the force per unit volume on the electricity moving with velocity $\boldsymbol{u}$ is $\rho[\boldsymbol{u}\cdot\boldsymbol{H}]/c$, the c being introduced because ρ is in electrostatic units. If we introduce the further idea that, in the presence of an electric field $\boldsymbol{E}$ there would be a force per unit volume equal to $\boldsymbol{E}\rho$,[21] in addition to $\rho\ [\boldsymbol{u}\cdot\boldsymbol{H}]/c$ the total force per unit charge would be

$$\text{Force per unit charge} = \boldsymbol{E} + \frac{[\boldsymbol{uH}]}{c}. \tag{101}$$

If we take a new unit of charge strength and a new unit of magnetic pole strength each respectively $1/(4\pi)^{\frac{1}{2}}$ of the electrostatic and magnetostatic units, and if the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ be defined as the forces on these units, the 4π disappears, as is well known from (97) and (99), and our equations become

$$\frac{1}{c}\left(\rho\boldsymbol{u} + \frac{\partial \boldsymbol{E}}{\partial t}\right) = \text{curl } \boldsymbol{H} \tag{102}$$

$$\rho = \text{div } \boldsymbol{E} \tag{103}$$

$$-\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t} = \text{curl } \boldsymbol{E} \tag{104}$$

$$0 = \text{div } \boldsymbol{H} \tag{105}$$

$$\text{Force per unit charge} = \boldsymbol{E} + \frac{[\boldsymbol{uH}]}{c}. \tag{106}$$

The addition of the equations $0 = \text{div } \boldsymbol{H}$ implies the additional postulate that all magnetic fields are traceable to currents without the invocation of any magnetic poles.

The foregoing arguments can hardly be regarded as a proof of the relations (102)–(106). As we have already remarked, the generalization involved in the addition of the term involving $\partial \boldsymbol{E}/\partial t$ in (102) could have been made in an infinite number of other ways as far as serving the purpose of providing for the equation of continuity is concerned. Again, the deduction of (104) as

[21] This is not a foregone conclusion, because even in the sense of the definition of $\boldsymbol{E}$ in terms of the force on a charge, that charge is supposed at rest.

applied to the field at a point in space from the (80) as applied to a complete material circuit is not an entirely rigorous procedure. Similar remarks may be made regarding (106). Moreover, the sense in which the term "force" is considered as applying in equation (106) is to some extent vague when considered in relation to distributions of electricity, for the scheme of equations itself provides no idea of electric mass to combine with acceleration as a criterion for a realization of the actions of the so-called force.

The best that can be said for the differential forms (102)–(104) in the light in which they have thus been obtained is that they are forms which when applied to the large scale phenomena of wire circuits, reproduce the equations (79) and (80). And even this statement is not correct except for steady currents for equation (79) contains no $\partial \boldsymbol{E}/\partial t$ term. It is quite true that the scheme represented by (79), (80), and (81), together with its moulding into a dynamical form as carried out by Maxwell represents a logical scheme which might have corresponded to nature even for varying currents in closed circuits. There are no inconsistencies in the scheme itself. The extension to the general case of moving charges has, however, necessitated the introduction of the $\partial \boldsymbol{E}/\partial t$ term; and, this term having been introduced, it must not be discarded on return to the simpler case of closed circuits unless it, of its own accord, disappears for that case, and this it does not do.

Maxwell was not satisfied with a mere deduction of (104) by the process we have sketched; but, having modified (95) by the introduction of the $\partial \boldsymbol{E}/\partial t$ term, he proceeded to seek an expression for the kinetic and potential energies per unit volume in terms of the electric fields and the current densities, which, by the application of the dynamical equations for continuous media would result in equations (102)–(106), in a manner analogous to that which he had adopted for the case of continuous circuits. His analysis suffered from certain imperfections, and it was later modified by Lorentz, Larmor, and others, with the special additional postulate that the electricity existed in the form of definite entities or electrons. The procedure amounted to choosing an expression for the kinetic and potential energies per unit volume of the medium as a function of the electric field and motion of the electricity, and then by means of the Hamiltonian dynamical principle $\partial \int (T - V) dt = 0$ applied to these energy expressions deducing two sets of equations, one of which was the equivalent of (104), and the other the equivalent of (106), but in a more unambiguous form. The form in question stated that the electron moves in such a manner that

$$\iiint \left(\boldsymbol{E} + \frac{[\boldsymbol{u}\boldsymbol{H}]}{c} \right) \rho d\tau = 0 \tag{107}$$

the integral being taken over the electron, and the fields representing the total fields, including the parts produced by the electron itself.

Again, these derivations can hardly be regarded as having accomplished more than the finding of the expressions which it is appropriate to call kinetic and potential energy in order that the customary dynamical procedure shall lead to the equations desired.

As a matter of fact, the form of the force equation symbolized in (107) is one which would be inconsistent with the principle of relativity; for, in the manner which it is presented by the dynamical development, it is of course not restricted to the case where the electron is momentarily at rest in the system of axes to which $\boldsymbol{E}$ and $\boldsymbol{H}$ are referred.

We shall presently return to another form of the expression of the dynamics of the electron, and one which is consistent with the theory of relativity; but for the moment, it will be well to make a digression to discuss a few matters concerned with the significance of (70)–(73), quite apart from the question of their origin.

The Properties and Dynamics of the Electron not Contained in the Circuital Relation.—In the first place it is to be remarked that while equations (70)–(73) may be solved so as to give the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ in terms of ρ and $\rho\boldsymbol{u}$ supposed assigned, they tell us nothing about the motions of the charges themselves. Since the equations are linear, we may add any two solutions and obtain as a result another solution. Thus, if we should find, as is in fact quite easy to do, the solution for the fields of an electron moving in a straight line with velocity v, and add to this the solution for another electron moving towards the first with velocity v, the result would also satisfy the equations. The equations themselves would be perfectly content for the electrons to go on in their motions quite oblivious of each other's presence. It is to the force equation entirely that we must look for an answer to how the motions of the electron influence each other. Had equations (70)–(73) been non-linear, this conclusion would not have held, solutions would not have been additive to give another solution; and, the equations themselves would have had something to say about the motions of the electrons relative to each other.

We frequently hear difficulties raised as to how the electron holds together, in view of the repulsion of its individual parts. Logically there is no difficulty. Equations (70)–(73) are perfectly content to have the various parts of an electron in mutual proximity without any tendency to fly apart. Even the force equation concerns only the motion of the electron as a whole; and it is no quibble to maintain that the ideas applicable to the whole entity are not applicable to its individual parts; for, as a matter of fact, even equation (107) while implying a motion of the electron which secures a zero value for $\iiint(\boldsymbol{E}+[\boldsymbol{uH}]/c)\rho d\tau$, by no means provides for the vector $\boldsymbol{E}+[\boldsymbol{uH}]/c$ being zero at each element of volume.

Solution of the Circuital Relations.—It is possible to solve equations (70)–(73) in such manner as to obtain the fields $\boldsymbol{E}$, $\boldsymbol{H}$, in terms of the charge density ρ, and the product $\rho\boldsymbol{u}$ supposed assigned. The results are:

$$\boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{U}}{\partial t} - \text{grad}\,\phi \tag{108}$$

$$\boldsymbol{H} = \text{curl}\,\boldsymbol{U} \tag{109}$$

where

$$4\pi c\boldsymbol{U} = \int\int\int \frac{(\rho \boldsymbol{u})}{r} d\tau \tag{110}$$

$$4\pi\phi = \int\int\int \frac{(\rho)}{r} d\tau \tag{111}$$

and where $\boldsymbol{U}$ and ϕ are still further subjected to the condition that

$$\text{div } \boldsymbol{U} = -\frac{1}{c}\frac{\partial \phi}{\partial t}. \tag{112}$$

The integrals are supposed taken throughout all space, and the parentheses are intended to denote that in finding the values of $\boldsymbol{U}$ and ϕ at time t, at any point (which is momentarily taken as the origin) we must insert in any particular element of volume the values of $\rho\boldsymbol{u}$, or ρ not at the time t, but at the earlier time $t-r/c$.[22]

Equations (70) and (71) of course contain the equation of continuity as a consequence. It can readily be shown that the relation (112) is the analytical guardian of the equation of continuity in preventing us from assigning in (110) and (111), values of $\rho\boldsymbol{u}$ and ρ which would of themselves violate that condition.

Thus, for example, if we were to assign $\boldsymbol{u}=0$, and $\rho=f(r)e^{-\alpha t}$ our assignment would of itself be inconsistent with

$$\text{div } \rho\boldsymbol{u} + \frac{\partial \rho}{\partial t} = 0$$

Equation (112) would prevent such an assignment of ρ and $\rho\boldsymbol{u}$.

The fact that the solutions represent ϕ and $\boldsymbol{U}$ and so $\boldsymbol{E}$ and $\boldsymbol{H}$ as the sum of a number of parts, each determined independently by the separate elements of charge enables us to speak of the field contributed by a charge A as distinct from the field contributed by some other charge B. Had, for example, products of the values of $\rho\boldsymbol{u}$ in the different elements of volume been involved in the expression for the field at a point, it would have been impossible to speak of a part of the field being due to one charge and part to another.

Definitions.—Before proceeding further, it is well that we pause for a moment to consider the significance which is to be attached to quantities like $\boldsymbol{E}$ and $\boldsymbol{H}$ in a scale of analysis which is contemplated in the equations as they are used in atomic physics. The definition of the magnetic field as the force on a unit magnetic pole, and of the electric field as the force on unit charge, are permissible so long as we are dealing with large scale phenomena; but, a unit pole has only existence in a macroscopic sense where a point is identified in a crude way with a region sufficiently large to contain many molecules. Even though, in the case of the charge, we can go to as

[22] We shall not pause to consider the possibilities inherent in the use of the advanced potentials.

small a unit as the electron, at any rate in imagination, for a charge entity in terms of which to define, by the magnitude of its acceleration, the electric field, our procedure becomes illogical when we contemplate a situation in which the field $\boldsymbol{E}$ varies enormously in magnitude and direction over the space comprised by a single electron. The question here involved is nor one of experimental difficulty, but of logical meaninglessness. It is not one purely ideal in its character, moreover; for in one of the most important applications of the equations, that involved in finding the quantity represented by the right-hand side of (75) we encounter a situation in which the quantity $\boldsymbol{E}$ with which we deal, varies to the extent of the whole of its value over the electron. Why is it that we are not led into hopeless error in cases of the above kind? It is because we do not *use* $\boldsymbol{E}$ and $\boldsymbol{H}$ in the sense of the definition in terms of unit charges and unit magnetic poles.

If we examine what we actually do in all problems of atomic dimensions we shall find that the starting point is the *assignment* of the charges and of their motions—the assignment of ρ and $\rho\boldsymbol{u}$. This is followed by the *calculation* of $\boldsymbol{E}$ and $\boldsymbol{H}$ by aid of equations (108)–(109), and the subsequent ultilization of the same $\boldsymbol{E}$ and $\boldsymbol{H}$ for some specific purpose, as, for example, the determination of the integral on the right-hand side of (75) (where only $\boldsymbol{E}$ is required, however).

From the standpoint which is the standpoint of their actual use in nearly all problems of atomic importance, equations (70)–(73) really constitute the definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$ in terms of ρ and $\rho\boldsymbol{u}$ supposed assigned.

But you may well ask what we gain by mere definition. We cannot discover laws of physics by definitions. That is true; but, we can, by well chosen definitions insure useful properties for the quantities defined. Thus, the quantities $\boldsymbol{E}$ and $\boldsymbol{H}$ defined by equations (70)–(73) vary with space and time in free space in accordance with the wave equation. Thus, for example:

$$\nabla^2\boldsymbol{E} - \frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2} = 0$$

and the wave equation insures for them certain other important properties—such, for example, as the property that in a plane wave the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ are perpendicular to each other, and lie in the plane of the wave-front. It would be useless to define things like $\boldsymbol{E}$ and $\boldsymbol{H}$ if we could not express in terms of them the behavior of something we can observe. The justification for their definition in electrodynamics lies in their being appropriate quantities in terms of which to express the motions of electrons through such equations as (75), or (78), for example; and, the properties provided for in their definitions provide in turn for properties which we can observe in the motions of the electrons referred to. Now it is quite true that, in the majority of problems of atomic dynamics, the things which happen to electrons do not go on in such manner as to indicate that they have been controlled by the classical force equation. Nevertheless, in formulating the new laws of such a phenomenon as the photoelectric effect, for example, the properties of $\boldsymbol{E}$ and $\boldsymbol{H}$ inherent in

their definitions through the electromagnetic equations make them useful props in a scaffolding appropriate to the new theory.

All that we need as a basis in terms of which to define a set of electromagnetic vectors satisfying equations (70)–(73) is a quantity ρ and a quantity $\rho\boldsymbol{u}$ satisfying the equation of continuity

$$\operatorname{div} \rho\boldsymbol{u} + \frac{\partial \rho}{\partial t} = 0$$

As a matter of fact, the density and density times velocity occur in the solutions for ϕ and $\boldsymbol{U}$, and so in those for $\boldsymbol{E}$ and $\boldsymbol{H}$ in the forms they do, not because of any property inherent in themselves, but simply on account of their relative positions in (70)–(73). If we should replace ρu_x, ρu_y, ρu_z, ρ by any four quantities S_x, S_y, S_z, R, satisfying

$$\operatorname{div} \boldsymbol{S} + \frac{\partial R}{\partial t} = 0$$

the solutions would be of exactly the same form as before, with $\boldsymbol{S}$ and R replacing $\rho\boldsymbol{u}$ and ρ. We shall return later to the significance of this aspect in relation to modern atomic theory.

Definitions in Terms of Assigned Entities.—For one to whom the aspect of (70)–(73) as definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$ is too artificial for his mental satisfaction, and who insists upon definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$ in terms of the supposed action of these vectors on something or other, a way is still open, at least to a scale of precision which regards the electron as no more than a point entity. It is only at the expense of considerable complication, however, that this way can be made precise.

The most drastic change from the classical procedure must be made in the matter of the magnetic pole. And, as is often the case, if we follow the plan of what actually happens in nature rather than invent artificial concepts, we gain in the long run. The action of a magnetic field on a pole is an idealization of its action on a magnet, and that is an idealization of the really more simple concept of its action on a moving electron, since the motions of electrons endow the magnet with its magnetic properties. It is then in the spirit of its action on a moving electron that we shall seek to define our magnetic field.

Let us suppose for a moment, that $\boldsymbol{E}$ and $\boldsymbol{H}$ had been defined in a manner which, for the moment, we need not specify, and let us suppose that the motion of an electron in terms of the fields so defined were given by (78), where we shall suppose the terms on the right-hand side other than the first to be absent. Then by putting $\boldsymbol{w}=0$, it would be a fact that

$$E_0 = \frac{m_0}{e} \operatorname*{Lt}_{w=0} \frac{d}{dt} \frac{\boldsymbol{w}}{(1 - \boldsymbol{w}^2/c^2)^{1/2}} \tag{113}$$

and that when $\boldsymbol{w}$ was not zero

$$\frac{[wH_0]}{c} = \frac{m_0}{e}\left\{\left(\frac{d}{dt}\frac{w}{(1-w^2/c^2)^{1/2}}\right)_{w\neq 0} - \operatorname*{Lt}_{w=0}\frac{d}{dt}\frac{w}{(1-w^2/c^2)^{1/2}}\right\} \quad (114)$$

If the direction of w be adjusted so that the quantity on the right-hand side is a maximum, we shall have

$$\frac{H_0}{c} = \frac{m_0}{|ew|}\left\{\left(\frac{d}{dt}\frac{w}{(1-w^2/c^2)^{1/2}}\right)_{w\neq 0} - \operatorname*{Lt}_{w=0}\frac{d}{dt}\frac{w}{(1-w^2/c^2)^{1/2}}\right\} \quad (115)$$

where the w inserted is chosen in the direction to make H_0 a maximum. The direction of H_0 is that consistent with the vector product of H_0 and the velocity w being in the direction of the vector specified as the right-hand side of (114).

Now the facts embodied in the above equations provide us with a means of defining E_0 and H_0. For, suppose, discarding the fact that E_0 and H_0 were supposed defined originally, we take (113) as a definition of E_0; and, after proceeding to the limit when $w=0$ in order to secure uniqueness, suppose we take (115) as the basis of our definition of H_0. That is:

$$\frac{H_0}{c} = \frac{m_0}{e}\operatorname*{Lt}_{=0}\frac{\left\{\left(\frac{d}{dt}\frac{w}{(1-w^2/c^2)^{1/2}}\right)_{w\neq 0} - \operatorname*{Lt}_{w=0}\frac{d}{dt}\frac{w}{(1-w^2/c^2)^{1/2}}\right\}}{|w|} \quad (116)$$

Then, the definitions so found will be consistent with the equation

$$e\left(E_0 + \frac{[wH_0]}{c}\right) = \frac{d}{dt}\frac{m_0 w}{(1-w^2/c^2)^{1/2}} \quad (117)$$

for all values of w. Of course, they would not *necessitate* that (117) should hold for all values of w.

If it did not turn out that (117) were true for all values of w when E_0 and H_0 were defined in terms of the limiting value of w, our definitions would be definite, at least, apart from further criticisms which I shall presently raise, but they would be useless. It would then be necessary to find in place of

$$\frac{d}{dt}\frac{m_0 w}{(1-w^2/c^2)}$$

a quantity W which, when we again made our definitions in a manner analogous to the above, gave us quantities E_0 and H_0, which were consistent with

$$e\left(E_0 + \frac{[wH_0]}{c}\right) = W \quad (118)$$

for all values of w.

Of course, it will be understood that our experimental realization of the above definitions would involve the measurement of such quantities as

$$\operatorname*{Lt}_{w=0}\frac{d}{dt}\frac{w}{(1-w^2/c^2)^{1/2}}$$

and would necessitate a more or less ideal experiment. In the case cited, we should have nothing but an acceleration to measure; but in general, this would not be the case. Thus, to take an example, suppose there existed certain quantities $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ in terms of which the equation of motion of the electron was

$$e\left(\boldsymbol{E}_0 + \frac{[w\boldsymbol{H}_0]}{c}\right) = a_0\ddot{\boldsymbol{S}} + a_1\dddot{\boldsymbol{S}} + a_2\ddddot{\boldsymbol{S}} + \cdots, \tag{119}$$

where $\ddot{\boldsymbol{S}}$ represents the acceleration, and the a's are functions of the velocity. On putting $w=0$, we obtain

$$e\boldsymbol{E}_0 = \bar{a}_0\ddot{\boldsymbol{S}} + \bar{a}_1\dddot{\boldsymbol{S}} + \bar{a}_2\ddddot{\boldsymbol{S}} + \cdots, \tag{120}$$

where the $\bar{a}$'s are the limiting values of the a's for zero velocity. We see that, in order to complete our definition it would be necessary to specify $\ddot{\boldsymbol{S}}$, $\bar{\ddot{\boldsymbol{S}}}$, . . . , etc., when $w=0$. Moreover, we cannot specify these arbitrarily. For example, we cannot say "consider an electron for which the velocity is zero and all the time derivatives of the motion are zero except $\ddot{\boldsymbol{S}}$," because, although $\ddot{\boldsymbol{S}}$ could adjust itself so that $e\boldsymbol{E}_0 = \bar{a}_0\ddot{\boldsymbol{S}}$ at the instant chosen, it would have to be in the process of change in order to make (119) fit the possibly new value of $\boldsymbol{E}_0$ and of $\boldsymbol{H}_0$ and w at the next and subsequent instants. Our only hope, in the case of such a situation as is represented in (120) would be, having made a satisfactory guess at the final form (119), to proceed to its limiting case (120), and then define $\boldsymbol{E}_0$ in terms of the actual values of $\ddot{\boldsymbol{S}}$, $\bar{\ddot{\boldsymbol{S}}}$, etc., as measured for that case. It is, of course, to be remarked that these actual values $\ddot{\boldsymbol{S}}$, $\bar{\ddot{\boldsymbol{S}}}$, $\bar{\bar{\ddot{\boldsymbol{S}}}}$, . . . , etc., would not be entirely unrelated. For, the specification of all the time derivatives of the motion at an instant determines that motion for all time if the function is analytic for all time; and, in general it determines the motion for some finite time if the function is analytic for a finite time. Now it is manifest that the motion so determined must be such that it agrees continually with (118) or with whatever equation happens to be the correct one; and, the left hand side of this equation involves the way in which the field $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ change with the time. We shall return later to consideration of some of these questions having to do with the motion of the electron.

Electrodynamics and Restricted Relativity. In the section dealing with restricted relativity, we have already called attention to the significance of the invariance of the circuital relation under the Lorentzian transformation and have remarked that before it is possible to consider the quantities $\boldsymbol{E}'$, $\boldsymbol{H}'$, defined by equations (20) and (21) as the true fields in the dashed systems, it is necessary to appeal to the definitions of these quantities and ascertain whether, in the sense of these definitions, $\boldsymbol{E}'$ and $\boldsymbol{H}'$ have, in the dashed system, a significance corresponding to that of $\boldsymbol{E}$ and $\boldsymbol{H}$ in the undashed system. Two cases are open for consideration. The first is that appli-

cable to the case where $\boldsymbol{E}$ and $\boldsymbol{H}$ are defined in terms of the motion of an assigned entity and the second is that applicable to the case where equations (70)–(73) constitute really the definitions of $\boldsymbol{E}$ and $\boldsymbol{H}$.

Significance of Relativity when $\boldsymbol{E}$ and $\boldsymbol{H}$ are defined in terms of the motion of an Assigned Entity. It is a matter of simple algebra to show that if $\boldsymbol{E}'$ and $\boldsymbol{H}'$ are related to $\boldsymbol{E}$ and $\boldsymbol{H}$ by equations (20) and (21), then if $k'=(1-w^2/c^2)^{-\frac{1}{2}}$ the four quantities $k'(\boldsymbol{E}'+w'\boldsymbol{H}'/c)$, $k'i(\boldsymbol{E}'w'/c)$ are related to the corresponding four quantities without dashes by the transformation of a 4-vector as specified in equations (53). So far there is no implication as to the meaning of $\boldsymbol{E}'$ and $\boldsymbol{H}'$, other than that defined by (20) and (21). Thus we know that if there be any such equation as (118), then provided that k times the three components of the right-hand side form three components of a 4-vector whose fourth component is $ki(\boldsymbol{W}w)/c$, the Lorentzian transformation when applied to this equation will lead to an exactly similar equation in dashed letters, with $\boldsymbol{E}'$ and $\boldsymbol{H}'$ having the significance implied in (20) and (21).[23]

Now since $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ are defined in terms of limiting cases of (118), the same definitions in terms of the corresponding limiting cases in the dashed system will insure that the quantities which we have called $\boldsymbol{E}_0'$ and $\boldsymbol{H}_0'$ will be the fields in the dashed system in terms of their proper definitions. Thus the correspondence between $\boldsymbol{E}_0$, $\boldsymbol{H}_0$, and $\boldsymbol{E}_0'$, $\boldsymbol{H}_0'$ is provided for, if there is an equation of motion of the type (118) where $\boldsymbol{W}$ has the properties specified above. If there were no such equation which was true for all velocities in the undashed system we might still use such an equation for the limiting cases to define $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$, as pointed out in the text between equations (117) and (118), but not only would the quantities thus defined be useless, but the corresponding equation when used for the purpose of providing corresponding definitions in the dashed system could not lead to quantities the equivalent of $\boldsymbol{E}_0'$ and $\boldsymbol{H}_0'$. If it did, the truth of the equation would be implied for finite velocities since the limiting cases in the dashed system would correspond to finite velocities in the undashed system.

Significance of Relativity when Equations (70)–(73) constitute the Definitions of E and H. Let us speak of the undashed and the dashed systems as S and S' respectively, and let us think of the elements of volume of

[23] Here we note in the first place that a factor k is necessary on each side of (118) to convert the quantities to the first three components of a 4-vector. This factor k may immediately be cancelled out however. In the second place, it will be noted that the fourth component of the right-hand side has to be $ki\ (\boldsymbol{W}.w)/c$. The point involved here is simply that we are not permitted to equate any two 4-vectors simply because they are 4-vectors, just as we are not permitted to equate any two ordinary vectors. The vectors must be in the same direction. The fact that the 4th component of the 4-vector must be $ki(\boldsymbol{W}.w)/c$ is bound up with a requirement analogous to that of equality of direction in three dimensions. There is no matter of fundamental interest involved here The question is one easily understood directly the mathematical technique is studied, and we shall not pursue it further. Finally, it may be remarked, we have talked of the quantity $\boldsymbol{W}$, rather than the more specific quantity $m_0/e\cdot d/dl\cdot w/(1-w^2/c^2)^{\frac{1}{2}}$ simply for generality. It may be well for the reader to think in the first instance of the more specific quantity referred to.

space in S as marked out with identifiable points at their boundaries, so that those boundaries may be recognized in S'. The elements of volume may be allowed to move about. Let A and A' be observers in S and S'.

Let us suppose that A and A' assign the same number ΔQ to corresponding elements of volume at times t and t' related by

$$t' = \beta\left(t - \frac{v}{c^2}x\right).$$

It is easy to show that corresponding elements of volume $\Delta\tau$ and $\Delta\tau'$ are related by the equation

$$\Delta\tau = \beta\left(1 - \frac{u_x v}{c^2}\right)\Delta\tau'$$

where u_x is the x component of the velocity of the element of volume in S.

If now we *define* ρ and ρ'' by[24]

$$\rho \equiv \frac{\Delta Q}{\Delta\tau}; \quad \rho'' \equiv \frac{\Delta Q}{\Delta\tau'} \tag{121}$$

it follows that

$$\rho'' = \beta\rho\left(1 - \frac{u_x v}{c^2}\right).$$

Thus the quantity ρ'' defined by (121) is the very same quantity as that defined by equation (22).

Suppose now A in S uses equations (70)–(73) or their solutions (108)–(112) to define $\boldsymbol{E}$ and $\boldsymbol{H}$ in accordance with the ideas outlined in the section on definitions. It will of course be true that the quantities $\boldsymbol{E}'$ and $\boldsymbol{H}'$ defined in terms of $\boldsymbol{E}$ and $\boldsymbol{H}$ by (20) and (21) will satisfy the equations obtained from (70)–(73) by dashing all the letters, and with the understanding that $\rho' = \beta\rho(1 - u_x v/c^2)$. If, however, the observer A' in S' defines what to him are the electric and magnetic fields he will arrive at those very fields $\boldsymbol{E}'$ and $\boldsymbol{H}'$ given by (20) and (21); for his charge density ρ'' has been proved equal to ρ', and the solutions of the electromagnetic equations in terms of charge density and velocity are unique insofar as we concern ourselves with the parts related to the charges.

Thus, the reality of the invariance, and the significance of $\boldsymbol{E}'$, $\boldsymbol{H}'$ and ρ' as true electric and magnetic fields and charge density depend ultimately upon the assignment of the same numbers by A and A' to the charges in corresponding elements of volume in the two systems.[25] But we must now

[24] We have used ρ'' instead of ρ' because ρ' has already been used in equation (22) as *defined* by $\rho' = \beta\rho(1 - u_x v/c^2)$, and we do not wish to assume that $\Delta Q/\Delta\tau'$ has this value until we have proved it.

[25] It is to be noted that although *invariance* of the charge is demanded, that is not the same thing as demanding *constancy* of the charge. As far as *invariance* was concerned the charge on an electron, for example, might vary with the time and all would be well, so long as A and A' measured equal charges in corresponding volume elements at corresponding times. It is true that the equations themselves provide for *constancy* of the charge as well, however.

inquire by what process A and A' arrive at conclusions as to the magnitudes of charges in corresponding volume elements. Of course they can only arrive at such conclusions by observing something which is observable, and which is calculable in terms of the fields. Thus, suppose having defined our fields as above, we invoke a force equation, for example, of the form (78). The magnitudes of the fields which we shall insert for $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ will depend upon the magnitudes which we assign to the charges in the elements of volume of space. It must be our business to assign the charges in such a manner that our equation of motion is true.[26] If we cannot make this equation of motion true, our business must be to find some other equation which is invariant under the transformation in the sense that (78) is invariant, and such that it will be possible to assign charges to the volume elements of space which will lead to the calculation of fields which make the equation true.[27] If we succeed in doing this in S, there is no further assumption involved in stating that in S' the corresponding charges will be the same. That they will be the same is a mathematical consequence of the forms of the equations. The dashed and the undashed system of equations are the expression of the same facts provided that the *corresponding charges in the two systems are chosen equal*; and, the process of fitting these equations to the same phenomena will insure that A and A' shall assign the corresponding charges as equal.

In order to crystallize the whole situation, even at the expense of some repetition, it may then be worth while to state the significance of invariance in relation to the electromagnetic equations. Our statement is as follows:

First we must imagine equations (70)–(73) supplemented by some such equation as (78) concerning which all we need say is that it describes the motion of some observable entity, and describes it in an equation which under the Lorentzian transformation assumes the same form in dashed letters, and in particular with $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ replaced by $\boldsymbol{E}_0'$ and $\boldsymbol{H}_0'$ as defined by, (20) –(21).

Then suppose that observer A uses equations (70)–(73) together with this extra equation to describe the motion of his observable entities. In this process, in order to calculate the fields $\boldsymbol{E}_0$, $\boldsymbol{H}_0$, he will have to assign certain charges to the volume elements of space (although of course, these charges will move about), and his charge densities must be supposed obtained by dividing these charges by the elements of volume which contain them. In terms of the charge densities and velocities assigned to the charges[28] he

[26] The constants m_0 and e in (78) combine to one constant m_0/e. All matters concerned with e being the same as the charges on the other electrons, or of m_0 being expressed in terms of e and of a radius of the electron arc matters perfectly consistent with, but not required by the considerations of invariance.

[27] Of course, there is no reason to expect that the charges so calculated will be grouped into equal units corresponding to electrons.

[28] It must be remembered that the velocity of an element of charge is not a quantity directly observable, and strictly speaking, these velocities will be determined as the numbers necessary to fit the facts, and, incidentlally, to preserve the equation of continuity $\partial\rho/\partial t + \text{div. } \rho\boldsymbol{u} = 0$, which equation is contained as a consequence of (70)–(73).

will calculate from (70)–(73) the fields E_0, H_0. He will then substitute the results so obtained in the additional equation to calculate the motions of the observable entity. The agreement of his calculations with the experimental facts is the criterion that he has assigned the correct charges and velocities; and, the fact that it is *possible* to find suitable charges and velocities to do this is the criterion that the theory is correct. Now the position of relativity in this matter is as follows: No extra assumption is involved. It is a mathematical consequence of the forms of the equations that if A' goes through exactly the same procedure to discuss the identical phenomenon which A is examining, he may use exactly the same equations. He need not even dash the letters, but it will of certainty result that if A was successful in calculating the observed motion, A' will be successful, and that moreover, the charges which A' will have to assign will be equal to those assigned by A, to corresponding elements at corresponding times. Of course, the fields which A' calculates will not be the same as those calculated by A. They will be related to them by equations (20) and (21). Moreover, the velocities will be related by (19) and the densities by (22). Of course, moreover, the form of the motion of the observable entity as calculated by A', in his coordinates, will differ from that calculated by A in his coordinates. It will be a motion obtainable from that calculated by A by the Lorentzian transformation. The point is that if the motion calculated by A agrees with A's measurements, the motion calculated by A' will agree with A''s measurements.

Perhaps one of the vital things to observe in all this is that the things which are assigned, the charges and the velocities, are either the same for the two observers as in the case of the charges, or are related by a law immediately predictable from the transformation itself, as in the case of the velocities. Had it been impossible to pass back from the relations between corresponding densities to the invariant charges, had it been necessary in fact, to say that A and A' must *assign* densities related by (22), much of the purity of meaning of the theory of relativity as applied to electrodynamics would have been lost.

The Relativity Dynamics of the Electron. Although the recent developments of the quantum theory have rendered it necessary for us to change our ideas as to the motion of an electron, and to discard the older force equations of electrodynamics as representative of the laws of motion in most atomic problems, nevertheless, the fact that classical laws do hold in limiting cases, of slowly varying forces, causes them still to form an important part in the dynamic story of atomic processes. In particular, the approximate equation

$$\left(E_0 + \frac{[wH_0]}{c}\right)e = m\frac{d}{dt}\frac{w}{(1-\epsilon^2)^{1/2}}$$

where $\epsilon = (w_x^2 + w_y^2 + w_z^2)^{\frac{1}{2}}/c$, still plays an interesting role, and it is customary to seek for it a parentage not so much in laws involving concepts of

the type represented by (107) as in laws as nearly as possible like those of classical dynamics.

According to the principles of general classical dynamics, the equation of motion of a system may be written in the form of Lagrange's equation

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_r}\right) - \frac{\partial L}{\partial q_r} = 0 . \tag{122}$$

Here no restriction is placed on L, other than the restriction that it is a function of the coordinates and the generalized velocities, and possibly also of the time. (122) is the analytical equivalent of

$$\delta \int L dt = 0. \tag{123}$$

where the values of δq_r at the time limits are zero. (122) contains as a consequence

$$\frac{d}{dt}\left\{ \sum \dot{q}_r \frac{\partial L}{\partial \dot{q}_r} - L \right\} + \frac{\partial L}{\partial t} = 0 . \tag{124}$$

If $L = T - V$ and T is a homogeneous quadratic function of the time rates of change of the coordinates, while V is a function of the coordinates and time, it is easy to show that

$$\sum \dot{q}_r \frac{\partial L}{\partial q_r} = 2T$$

and equation (124) becomes

$$\frac{d}{dt}(T + V) + \frac{\partial L}{\partial t} = 0$$

so that if V does not involve the time explicitly

$$\frac{d}{dt}(T + V) = 0$$

and

$$T + V = \text{constant}$$

$T + V$ is of course the energy. Even where L is not of the above form, however, it is still true that if L does not involve t explicitly, $\Sigma \dot{q}_r(\partial L/\partial \dot{q}_r) - L$ is constant with time, so it is convenient to give the name "energy" to this quantity for the general case

$$\text{Energy} = \sum \dot{q}_r \frac{\partial L}{\partial q_r} - L$$

and understand by this the energy, even for cases where L does involve t explicitly. An analytical equivalent of (122) or (123) is

$$\dot{q}_r = \frac{\partial H}{\partial p_r}; \quad \dot{p}_r = -\frac{\partial H}{\partial p_r} \tag{125}$$

where p_r is defined as

$$p_r \equiv \frac{\partial L}{\partial \dot{q}_r} \tag{126}$$

and H is defined as

$$H \equiv \sum p_r \dot{q}_r - L \tag{127}$$

where, however, the $\dot{q}$'s in H are expressed in terms of the q's and p's through equation (126), so that H is a function of the p's and q's and of t.

It is an analytical consequence of (125) that

$$\frac{dH}{dt} = -\left(\frac{\partial L}{\partial t}\right)_{p\text{'s and } q\text{'s constant}} \tag{128}$$

Again, if $L = T - V$ and T is a homogeneous quadratic function of the velocities and V a function of coordinates and time, $H = T + V$, and is constant with time if L does not involve t explicitly.

In general, if

$$L = T' - V,$$

where T' is merely a function of velocities, coordinates and time, but not a homogeneous quadratic function, the energy $\Sigma p_r \dot{q}_r - L$ will be something other than $T' + V$, and it will be this something which will be constant with time or equal to $-(\partial L/\partial t)$ according as L does or does not involve t explicitly.

Let us consider the consequences of taking for L,

$$L = \sum_i - m_{0i}c^2(1 - \epsilon_i^2)^{1/2} - e\phi_i + \frac{e}{c}(U_{xi}w_{xi} + U_{yi}w_{yi} + U_{zi}w_{zi}) \tag{129}$$

where

$$c^2\epsilon_i^2 = w_{xi}^2 + w_{yi}^2 + w_{zi}^2$$

where U_i and ϕ_i are the vector and scalar potentials contributed by the charges other than that of the ith electron. The quantity L satisfies the requirements we have imposed on L, the requirements of being a function of the q's, the $\dot{q}$'s and Equation (126) then leads to

$$f_i = \frac{m_{0i}w_i}{(1 - \epsilon_i^2)^{1/2}} + \frac{e}{c}U_i. \tag{130}$$

On making use of this in (127) we find

$$H \equiv \sum_r p_r \dot{q}_r - L = \sum_i \left\{ \frac{m_{0i}c^2}{(1 - \epsilon_i^2)^{1/2}} + e\phi_i \right\} \tag{131}$$

so that if ϕ_i and U_i are independent of t explicitly, so that L is independent of t explicitly

$$\sum_i \left\{ \frac{m_{0i}c^2}{(1 - \epsilon_i^2)^{1/2}} + e\phi_i \right\} = \text{constant}.$$

Whether this quantity is constant with time or not, it is, as we have remarked what we call the energy, since it is the quantity which is constant when L is independent of t explicitly.

In the case of a single electron[29] and an external field specified by ϕ_i and U_i, as functions of position and time, ϕ_i involves only the coordinates and time, so that it is appropriate to call $e\phi_i$ the potential energy. It is then appropriate to call $m_{0i}c^2/(1-\epsilon_i^2)^{1/2}$ the kinetic energy T. Or since it is customary to regard T as zero when ϵ is zero, we take

$$T \equiv m_{0i}c^2\left(\frac{1}{(1 - \epsilon^2{}_i)^{1/2}} - 1\right).$$

The addition of the constant term does not, of course, affect the condition

$$\frac{d}{dt}(T + V) = -\frac{\partial L}{dt}.$$

The quantity we have called T' is, in the present case,

$$T' = -m_{0i}c^2(1 - \epsilon_i^2)^{1/2} + \frac{e}{c}\left(U_{xi}w_{xi} + U_{yi}w_{yi} + U_{zi}w_{zi} \right).$$

Whether ϕ_i and U_i are, or are not, independent of t explicitly, it results on applying (122) or (123) to L, with due regard to the proper restricting conditions under which the variation is performed, that

$$\left\{ -\frac{1}{c}\frac{\partial U_i}{\partial t} - \text{grad}\,\phi_i + \frac{1}{c}[w_i \text{ curl } U_i] \right\} e = \frac{d}{dt}\left(\frac{m_{0i}w_i}{(1 - \epsilon_i^2)^{1/2}} \right) \tag{132}$$

so that from (108) and (109)

$$\left(E_{0i} + \frac{[w_i H_{0i}]}{c} \right) e = m_{0i}\frac{d}{dt}\left\{ \frac{w_i}{(1 - \epsilon_i^2)^{1/2}} \right\}. \tag{133}$$

Thus we may regard (129) as the appropriate expression for L in order to give rise to the equations of motion (133) by the usual dynamical procedure. In (132) we should regard U_i and ϕ_i as defined in terms of the charges of

[29] Some little care is necessary in interpreting the exact significance of this matter for the case of more than one electron, since the retarded nature of U and ϕ cause them to involve higher time derivatives of the motion other than the velocity. It is, however, possible to regard U and ϕ as functions of the time explicitly, and then take account of their relation to the electronic motions through the introduction of undetermined multiples in the usual manner when L is subjected to the variational operation. It is not my intention here to do more than call attention to the point as a complete discussion of it would involve more space than is consistent with its importance.

the *external* system by (110) and (111). The quantities E_{0i} and H_{0i} need never be used at all except in the light of a shorthand notation. If they are introduced as a matter of convenience, they must be regarded in the light of being defined by (108) and (109); and, in the light of this definition, their obedience to (70)–(73) is guaranteed. We have to choose our Lagrangian function L in any case; and, having chosen it in terms of ϕ and U, the complete story is told in a way which is definite. It may be wrong, but it is definite, and E_{0i} and H_{0i} need never make their appearance at all except in the character of a scaffolding. It is of interest to note that the quantity which we usually call the momentum of the ith electron, viz.: $m_{0i}w_i/(1-\epsilon_i^2)^{1/2}$ is not the same thing as the generalized momentum of the system *associated* with the electron i, the latter quantity is given by (130).

It is perhaps worthwhile to emphasize that the application of (122) or (123) to (129) to obtain (132) or (133) is not to be regarded as a *proof of* (132) or (133), since the choice of (129) rests on its being the suitable function for this purpose. All that is shown is that (132) or (133) is the analytical equivalent of applying (122) or (123) to (129). Indeed, it is probable that most so-called proofs in physics should be regarded more in the light of showing the analytical equivalence of different statements of a law.

We know that, with suitable definitions of the quantities concerned as already discussed, (133) for a single electron is invariant under the restricted relativity transformation. We should expect that this would be the case if Ldt for a single electron were invariant in form under the Lorentzian transformation. It is easy to see that this invariance holds. For, writing $k_i \equiv (1-\epsilon_i^2)^{-1/2}$, dt/k_i is known to be an invariant. Also, $k_{xi}w_{xi}$, $k_{yi}w_{yi}$, $k_{zi}w_{zi}$, $kc\sqrt{-1}$ is a 4-vector. It is the velocity 4-vector. Also U_{xi}, U_{yi}, U_{zi}, $\phi\sqrt{-1}$ is a 4-vector. Thus $m_{oi}c^2(1-\epsilon_i^2)^{1/2}dt = m_oc^2dt/k_i$ and is invariant. Also $U_{xi}w_{xi}+U_{yi}w_{yi}+U_{zi}w_{zi}-c\phi_i = (U_{xi}k_iw_{xi}+U_{yi}k_iw_{yi}+U_{zi}k_iw_{zi} - ck_i\phi_i)dt/k_i$ and is therefore invariant. Hence, Ldt is invariant.

It is, of course, possible to deduce (133) from the Hamiltonian equations (125). For this purpose we need the Hamiltonian function for an electron. This is to be obtained by substituting for the $\dot{q}$'s in (127), their values in terms of the q's and p's from (130). In this way we readily obtain for a single electron

$$H = c\left[\left(p_x - \frac{e}{c}U_x\right)^2 + \left(p_y - \frac{e}{c}U_y\right)^2 + \left(p_z - \frac{e}{c}U_z\right)^2 + m_0^2c^2\right]^{1/2} + e\phi. \quad (134)$$

Intuitive Concepts on Electrodynamics: In the historical developments of electrodynamics, attempts to visualize the processes going on in the field have lead to certain forms of statement of the content of the equations which are applicable only in certain cases, but which have in certain minds become so fundamentally engrained as to assume a priority in respect to the equations themselves. Prominent among these concepts is that of

moving lines of magnetic force.[30] It is customary to say that a moving magnetic field produces an electric field perpendicular to the lines of force and to the direction of motion of the lines of force, the magnitude of the field being proportional to the product of the velocity and the magnetic induction. It is easy to invent cases where this statement leads obviously to error. Thus imagine two cylindrical bar magnets with axes in line, separated by a gap, so that the north end of one magnet faces the south end of the other. It would be natural to think of the rotation of this system about its axes as giving rise to a squirrel cage-like motion of the lines of magnetic force; and one who supposed that there was an electric field perpendicular to the lines of force and to the velocity at each point would have to conclude that the electric field was radial, and he would be lead to the unwelcome conclusion that the lines of electric force must converge to a linear charge on the axis of the system between the poles. A direct appeal to the electromagnetic

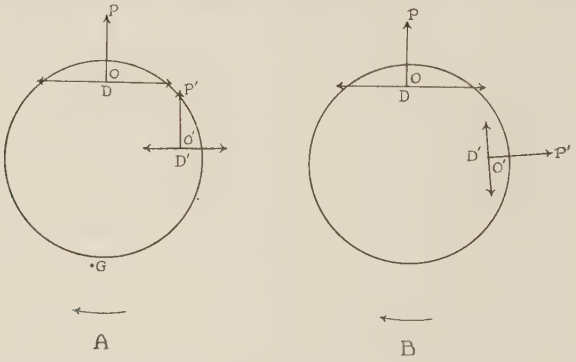


Fig. 3.

equations would, of course, deny the existence of such an electric field. But, while it is easy to construct cases of this kind, and show that the fallacious conclusions would not be predicted by a proper application of the electromagnetic equations, one must face the conclusion that in certain important and simple cases, such as the bodily translatory movement of a magnet in the vicinity of a wire, the law of the rate of cutting of lines of force does lead to a correct calculation of the electromotive force in the wire; and, an exponent of the moving line idea may well claim that there must be something in his statements.

Let us imagine ourselves facing the end of a cylindrical magnet represented by the large circles in Fig. 3, *A* and *B*. Let us further suppose that the magnet is in rotation in the sense indicated by the curved arrow. Let us moreover, think of the individual molecular magnets of which the large magnet is composed. In the first place, it is easy to see that there is an am-

[30] A fairly full discussion of the whole subject concerned with moving lines of force will be found in a paper "Unipolar Induction," by W. F. G. Swann, Phys. Rev., Vol. XV, No. 5, pp. 365–398, May, 1920; also in "Unipolar Induction" by John T. Tate, a section from "Electrodynamics of Moving Media," Bulletin No. 24, of the National Research Council, Part II, pp. 75–95.

biguity in meaning as regards the motion of the lines of force of the resultant field and the motions of the lines of force of the individual molecular magnets. Thus, the resultant lines of force of the magnet at such a point as G are moving from right to left, whereas the contribution to the lines of force at G from the molecular magnet D is such as to correspond to a motion from left to right. Thinking for a moment of the molecular magnets and representing by the radial arrows drawn at D and D', the lines of force coming out of a pole of the molecular magnet, there are two extreme possibilities as regards the meanings to be attached to the motion of the line of force OP for example. The motion may be thought of as corresponding to Fig. 3, A, in which in the motion of the molecular magnet from D to D', the line of force remains parallel to itself. Or, it may be thought of as corresponding to Fig. 3, B, in which, in the motion from D to D', the line of force OP assumes the position there shown. It can be shown that it is by thinking of the motion of lines of force in the former sense that we reproduce the equivalent of the electromagnetic equations. But, in this picture, it is impossible to combine the lines from the different molecular magnets into a resultant, and use the motion of that resultant as a basis for the discussion. On the other hand, in the latter alternative, as represented at Fig. 3, B, it is easy to see that the story predicted by the combined effect of all the molecular magnets would be the same as that predicted by combining the fields into a resultant which rotated rigidly with the magnet as a whole. The attachment of this significance to the motion of lines of force leads, however, to results totally inconsistent with those of the electromagnetic equations.

In the case of the pure translatory motions of magnets without rotation, the former case is the only one which can arise so that in such cases the moving line of force idea fits in with the facts.

The Position of Electrodynamics in Modern Atomic Theory, in Wave-Mechanics, etc. This is not the place to discuss the origin and content of Wave mechanics, and we must be content to dissect out for discussion simply those parts in which electrodynamics figure; treating the remainder as mere statements of methods of procedure. Even here, it will be necessary to confine attention in the present report to a very limited phase of the subject, and omit entirely the electrodynamics theory of Dirac, since to include it would lead us too far into subjects with which it is not the primary purpose of this article to deal.

It will then be recalled that in discussing a problem in wave mechanics, the procedure is first to write down the Hamiltonian function H which would have figured had the intention been to discuss the problem on the lines of classical dynamics. H is a function of, say, n coordinates q_r and n momenta p_r, and of the time t

$$H = H(q_1, q_2 \cdots q_n, p_1, p_2, \cdots t) \tag{135}$$

The next thing is to construct for a quantity ψ a differential equation

$$-\epsilon\frac{\partial\psi}{\partial t} = H\left(q, q\cdots\epsilon\frac{\partial}{\partial q}, \epsilon\frac{\partial}{\partial q_2}\cdots t\right)\psi \tag{136}$$

where $\epsilon = h/2\pi i$, and p_r in (135) is replaced by $\epsilon(\partial/\partial q_r)$ in the operator $H(q_1, q_2\cdots\epsilon\partial/\partial q_1, \cdots, t)$ which operates on ψ in (136). The next procedure must be to solve this differential equation subject to the boundary conditions appropriate to the problem, and when the solution is found to construct $\bar{\psi}$ the conjugate to ψ. The content of the theory lies in the supposition that if we consider any entity of the system specified by coordinates $q_s, q_{s+1}, \ldots . q_{s+p}$, the probability that these coordinates lie between q_s, $q_{s+1}\cdots q_{s+p}$, and q_s+dq_s, $q_{s+1}+dq_{s+1}, \cdots, q_{s+p}+dq_{s+p}$ is

$$\delta P = dq_s dq_{s+1}\cdots dq_{s+p}\cdots\int\cdots\int\psi\bar{\psi}(dq_1dq_2\cdots dq_{s-1})(dq_1\cdots dq_{s1})\,(dq_{s+p+1}\cdots dq_n) \tag{137}$$

Incidentally, it may be remarked that the functions $\psi\bar{\psi}$ are supposed normalized so that $\int\cdots\int\psi\bar{\psi}\,dq_1dq_2\cdots dq_{n=1}$.

Everything depends, of course, upon the choice of the Hamiltonian function; and, in the developments of the theories of various phenomena, it cannot be said that wave mechanics has been made to occupy a very consistent position. It will be sufficient, in the first instance to confine ourselves to a case of three dimensions i.e. the case of an electron in a prescribed field of force.

The first problem to engage our attention is naturally that of line spectra. Here the procedure has been to write down the simple Hamiltonian function for the atom in the Newtonian form, with no terms depending upon electrodynamics. The determination of ψ then follows from the solution of (136) subject to the usual conditions of vanishing at infinity,[31] and it corresponds to the sum of a number of terms simple harmonic in t. An expression for $\psi\psi$ is then obtained, and this expression changes with time at any point in a manner representative of the sum of a number of simple harmonic periods corresponding to those of the line spectra of the atom. For the usual case, that where (136) involves the p's as the sum of squares, it turns out as a direct consequence of that equation that the integral of $\psi\bar{\psi}$ throughout all space is independent of the time.

For the usual case, viz. that where (136) involves the p's as the sum of squares, of the form $(p_x^2+p_y^2+p_z^2)/2m$ it turns out as a direct consequence of that equation that[32]

$$\frac{\partial\psi\bar{\psi}}{\partial t} = \frac{h}{4\pi im}\,\mathrm{div}\,[\bar{\psi}\,\mathrm{grad}\,\psi - \psi\,\mathrm{grad}\,\bar{\psi}] \tag{138}$$

If we define a quantity w by the equation

[31] A condition which places restrictions upon the values of the possible constants of integration of (136), a restriction limiting them to the well-known Eigenwerte.

[32] Strictly speaking what results is that the space integrals of the left and right handsides of this expression are equal. Equation (138) however provides at any rate a sufficient condition for the conservation of the space integrals of $\psi\bar{\psi}$.

$$\psi\bar{\psi}\cdot\boldsymbol{w} = \frac{h}{4\pi i m}\,\mathrm{div}\,[\bar{\psi}\,\mathrm{grad}\,\psi - \psi\,\mathrm{grad}\,\bar{\psi}] \tag{139}$$

equation (138) may be regarded as an equation of continuity with $\psi\bar{\psi}$ functioning in the position of a density and $\boldsymbol{w}$ in the position of a velocity.

The equation of continuity having been provided for, the quantities $\psi\bar{\psi}$ and $\psi\bar{\psi}\cdot\boldsymbol{w}$ provide suitable quantities in terms of which to define a set of vectors $\boldsymbol{E}$ and $\boldsymbol{H}$, satisfying equations of the type (70)–(73), in terms of which the action of radiation on some other atom may be discussed.

By proceeding to the limiting case for a doublet of moment $\boldsymbol{M}$ in electromagnetic theory, we find that ϕ and $\boldsymbol{U}$, and so $\boldsymbol{E}$ and $\boldsymbol{H}$ are completely determined by the assignment of $\boldsymbol{M}$ and $\dot{\boldsymbol{M}}$, so that any quantity $\boldsymbol{M}$ varying with time and assigned to a particular point of space may be used as a basis for defining a set of electromagnetic vectors conforming to the frequencies contained in the time variations in $\boldsymbol{M}$. The procedure for obtaining, for example, the x component of the moment, is to evaluate $\int\int\int\psi\psi x dt$ integrated throughout all space, and take this quantity for $\boldsymbol{M}$.

The vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ having been provided for in some such manner as the foregoing, we may then proceed to calculate their values at other points in space where we wish to discuss some phenomenon such as the Compton effect, for example.

The electromagnetic fields defined in this way are, of course, not ends in themselves; and, in discussing such a phenomenon as the photoelectric effect, for example, produced by these fields, the procedure is to write down for an electron in the atom a Hamiltonian function of the form (134) with the quantities ϕ and $\boldsymbol{U}$ as the scalar and vector potentials corresponding to the electromagnetic radiation as above defined. The subsequent procedure is to solve the equation of the type evolved from this Hamiltonian function and the value of $\psi\bar{\psi}$ which results from this solution provides for the history of the ejected photoelectrons by giving the probability that an electron may subsequently be found in any assigned region of space.

A slightly different type of problem is represented by the scattering of x-rays by an electron. Here the customary starting point is the assignment of a Hamiltonian function of the type (134) for the electron, $\boldsymbol{U}$ being now the vector potential of the primary x-rays and ϕ the total potential due to x-rays and atom. The solution of the corresponding equation of the type (136) then follows, subject to appropriate boundary conditions, and the value of $\psi\bar{\psi}$ is calculated at each point of space. The secondary, or scattered radiation, is then calculated as the electromagnetic radiation emitted with $\psi\bar{\psi}$ regarded as the charge density.[33] Strictly speaking, the detection of this radiation would involve the observation of such a phenomenon as the subsequent emission of a photoelectron by the light, and the story in this connection would have to be told in manner analogous to that indicated above for the detection of the light emitted in line spectra.

[33] In the treatment by W. Gordon, (Zeits. f. Physik **40**, 117, 1926), the charge density is taken as a quantity somewhat different from $\psi\bar{\psi}$, but this consideration is not pertinent to the present discussion.

The most unsatisfactory part of the situation concerned in such problems as the foregoing lies in the association of $\psi\bar{\psi}$ with the charge density. Sometimes an effort is made towards a reconciliation of the situation with the logical demands by saying that since $\psi\bar{\psi}$ at a point represents the probability of the presence of an electron in unit volume at that point, the time changes of $\psi\bar{\psi}$ in a volume element represent on the average the time changes of concentration of actual electrons in the volume element. Such a view ignores the fact that, in calculating the fields from the $\psi\bar{\psi}$ distribution, the phase relations of the time oscillations of $\psi\bar{\psi}$ at different places play an important part; and, such phase relations would not necessarily be preserved in the time rates of change of concentration of real electrons if $\psi\bar{\psi}$ represents only the *probability* of the presence of an electron per unit volume. The situation is analogous to that in which a person has two flames of equal intensity, and hopes to get interference between them, forgetful of the fact that the absence of phase relations between the flames would destroy the possibility of occurrence of such phenomena. It is, however, possible to discard entirely the idea that the real electrons have anything to do with the production of the electromagnetic fields associated with the potentials ϕ and U, and to regard these quantities as determined entirely by the quantity $\psi\bar{\psi}$. The procedure amounts to actually understanding by the vector and scalar potentials occurring in the original equation for ψ not scalar and vector potentials calculated in terms of the actual electrons, but quantities defined by

$$\left.\begin{aligned} 4\pi\phi &= \int\int\int \frac{(\psi\bar{\psi})d\tau}{r} \\ 4\pi c U &= \int\int\int \frac{(\psi\bar{\psi}\cdot w)}{r} d\tau \end{aligned}\right\} \tag{140}$$

with w defined by (139) and with the potentials taken in the retarded sense.

A word may be added concerning problems involving many dimensions. Returning to expression (137) for the probability, and considering the entities concerned as n electrons, the probability ΔP_1 that electron No. 1 lies in the range dq_1 regardless of the positions of the others is

$$\Delta P_1 = dq_1 \int\int\int \cdots \int \psi\bar{\psi} dq'$$

where here, by dq_1, we shall understand the product $(dq_1)_x(dq_1)_y(dq_1)_z$ associated with the three coordinates of electron No. 1, and where dq' refers to the product of all the differentials other than $(dq_1)_x$, $(dq_1)_y$, and $(dq_1)_z$. Similarly the probability ΔP_2 that electron No. 2 lies in the range dq_2 regardless of the positions of the others is

$$\Delta P_2 = dq_2 \int\int \cdots \int \psi\bar{\psi} dq''.$$

A similar result holds for all of the electrons. Thus, the probability ΔP that

some unspecified electron lies in the element $dx\ dy\ dz$ of three dimensional space, regardless of the positions of the other electrons is

$$\Delta P = \sum_{\mu} dq_{\mu} \int\int \int \cdots \int \psi\bar{\psi} dq_{\mu}^{*}$$

where dq_{μ}^{*} refers to the product of all of the differentials other than dq_{μ}, where the sum is taken for all the coordinate elements dq_{μ} arising from the electrons, and where ultimately, all the dq_{μ}'s outside the integrals are replaced by $dx\ dy\ dz$, so that

$$\Delta P = dxdydz \sum f_{\mu}(x \cdot y \cdot z) \tag{141}$$

Here $f_{\mu}(x \cdot y \cdot z)$ is a function of $x \cdot y \cdot z$ in ordinary three dimensional spaces.

Thus, in any many dimensional problem it is this quantity which may function as a charge density for building up vector and scalar potentials to introduce into the Hamiltonian function. With it must be associated a velocity so chosen as to satisfy the equation of continuity for the apparent charge density. As a matter of fact in order to provide consistency with a case in which the electrons associated with the problem may be supposed to have different charges, it is desirable to associate with each component of the sum in (141) a separate velocity w_{μ}, defined by (139) with the $\psi\bar{\psi}$ of that expression replaced by $f_{\mu}(x \cdot y \cdot z)$.

In line with equation (134), the natural Hamiltonian function to assume for a group of electrons would be

$$H = \sum c \left[\left(p_{xi} - \frac{e_i}{c} U_{xi} \right)^2 + \left(p_{yi} - \frac{e_i}{c} U_{yi} \right)^2 + \left(p_{zi} - \frac{e_i}{c} U_{zi} \right)^2 + m_0^2 c^2 \right]^{1/2} + \sum e_i \phi_i \tag{142}$$

and, writing $\epsilon = h/2\pi i$, the corresponding equation for ψ would be

$$-\epsilon \frac{\partial \psi}{\partial t} = \left\{ \sum_i c \left[\left(\epsilon \frac{\partial}{\partial x_i} - \frac{e_i}{c} U_{xi} \right)^2 + \left(\epsilon \frac{\partial}{\partial y_i} - \frac{e_i}{c} U_{yi} \right)^2 + \left(\epsilon \frac{\partial}{\partial z_i} - \frac{e_i}{c} U_{zi} \right)^2 + m_0^2 c^2 \right]^{1/2} + \sum e_i \phi_i \right\} \psi \tag{143}$$

where

$$\left.\begin{aligned} 4\pi\phi_i &= \sum \int\int\int \frac{\{f_{\mu}(x \cdot y \cdot z)\}}{r} d\tau \\ 4\pi c U_i &= \sum \int\int\int \frac{\{w_{\mu} f_{\mu}(x \cdot y \cdot z)\}}{r} d\tau \end{aligned}\right\} \tag{144}$$

where the integrals are retarded in the usual manner, and where the summation in the case of ϕ_i and U_i is taken for all the f_{μ} other than that originating on account of the i^{th} electron.

A ψ equation based upon the Hamiltonian function above given has been used by W. Gordon,[34] but without the expression for ϕ_i and U_i in terms of the f_μ as above. The use of a ψ equation based upon this Hamiltonian function has the disadvantage of requiring, for the strict applicability of the equation of continuity, a definition of equivalent charge density and charge velocity other than those which we have given, and as a matter of fact this inconsistency exists even in equations (142)–(144). It may be avoided in a formal manner by holding to our definition of equivalent charge density constituted by each electron and associating with it a velocity w_μ defined by[35]

$$w_\mu \cdot f_\mu(x \cdot y \cdot z) = w\mu \frac{1}{4\pi} \operatorname{grad} \int\int\int \frac{\frac{\partial}{\partial t}\{f_\mu(x \cdot y \cdot z)\} d\tau}{r}$$

which, since the divergence of the right-hand side is $-\partial/\partial t \{f_\mu(x \cdot y \cdot z)\}$, leads to an equation of continuity

$$\operatorname{div} \{f_\mu(x \cdot y \cdot z)\} + \frac{\partial}{\partial t}\{f_\mu(x \cdot y \cdot z)\} . = 0$$

It will be recalled that in our former discussion, the invariance of the electromagnetic equations implied the assignment of equal charges by the observers A and A' to corresponding elements at corresponding times in the two systems. In the present formulation, this becomes translated into the assignment of equal values of $f_\mu(x, y \cdot z) d\tau$; in other words, of equal probability of an electron being found in corresponding volume elements at corresponding times. This condition is a very natural one, and is indeed demanded by the concept of such an entity as an electron, which is a thing primarily supposed as recognizable by one observer if it is recognizable by the other. It is a comparatively easy matter to show that, for the three dimensional case with, of course ϕ_i and U_i given as explicit functions of the time, (143) is invariant under the Lorentzian transformation. For the case of many dimensions, however, this result does not hold.

In conclusion, the reader must again be warned against regarding the foregoing as anything like an adequate treatment of electrodynamics in relation to modern wave mechanics. The excuse for the relatively fragmentary treatment given is that it perhaps clarifies certain of the major points which may have caused difficulty in the minds of those whose interest in the subject has been more or less that of the general reader.

[34] loc. cit.

[35] The integral in this expression is, of course, not retarded

THE APPLICATION OF GROUP THEORY TO THE QUANTUM DYNAMICS OF MONATOMIC SYSTEMS

By Carl Eckart
University of Chicago

Table of Contents

I. General Theory

1. The wave equation and the manifold of its solutions..... 306
2. The integrals of the wave equation and the finite matrices representing them..... 309
3. Further restriction of the operators to be considered. The group postulates..... 311
4. Invariant manifolds..... 313
5. Schur's lemmas..... 314
6. Application to perturbation theory..... 317
7. The multiplication table of a group. The rearrangement theorem..... 322
8. The orthogonality relations..... 326
9. The complete reduction of a given invariant manifold..... 328
10. The theory of characters..... 329
11. Miscellaneous theorems..... 331

II. Specification of the Groups to be Considered

12. The integrals in the absence of an external field. The case of a single electron..... 333
13. Extension to the case of N electrons..... 337
14. The integrals in the presence of an external field..... 337

III. Construction of Irreducible Representations of the Rotation Group, etc.

15. The group I, K..... 338
16. The two-dimensional rotations..... 339
17. The groups I, S, K', SK', and I, S, K'', SK''..... 340
18. The unitary substitutions of two variables $D(\alpha\beta\gamma)$..... 341
19. The three dimensional rotations..... 344

IV. Application to the Wave Equation

20. The wave equation without spin terms..... 347
21. The infinitesimal rotations and the angular momentum operators..... 349
22. The vectorial addition of angular momenta as the reduction of an invariant manifold..... 351
23. The method of sums..... 356
24. The selection principles..... 357
25. The addition of spin-vectors as the reduction of an invariant manifold. The alternation rule for even and odd multiplets. 360

V. The Permutation Group and the Pauli Exclusion Principle

26. Physical considerations regarding the necessity and possibility of an exclusion principle. 364
27. The permutation group and its representations. 365
28. The Pauli exclusion principle and the structure of normal multiplets. 366
29. The enumeration of the multiplets arising from a given electronic configuration. 370
30. Evaluation of the perturbation matrix. 375

VI. Influence of External Fields

31. The Zeeman effect. The sum rules for the g-factors. 378
32. Stark effect. 380

THIS report is intended to furnish an account of so much of the theory of groups and their representations as has been found useful in the solution of quantum dynamical problems. The general theory (Part I) is general only in comparison to the applications which follow. This is evident in the treatment of Schur's lemma, which is proven for a very special case, mathematically speaking, but with quite sufficient generality for the purposes of quantum dynamics.

The greater portion of the report is taken up by applications to the theory of complex atomic spectra. It has seemed better to limit the field of application and to use the available space for a fairly detailed account of some problems, rather than to draw a few general illustrations from each of several fields. This may lead the reader to form a false estimate of the applicability of the methods; it may merely be mentioned that they have also been applied to the theories of molecular spectra, valence forces, electron conduction, and ferromagnetism.

I. General Theory

1. **The wave equation and the manifold of its solutions.** The most important computational task set by quantum dynamics is the solution of the characteristic value problem

$$H\psi = W\psi. \tag{1}$$

Here ψ is a function of the translatory and spin coordinates of the electrons: these will be denoted by $x = (x_1 x_2 \cdots s_1 s_2 \cdots)$; H is a linear operator, and W a number independent of x. Certain other conditions (the "natural" boundary conditions) are imposed on ψ; those solutions of Eq. (1) which satisfy these are called the characteristic solutions of the equation, and those values of W for which at least one such solution exists are called the characteristic values or energy levels. This problem has been discussed in its relationship to the physical principles of the quantum theory elsewhere in this journal;[1] in this review it is proposed to discuss the problem

[1] E. C. Kemble, Phys. Rev. Supplement 1, 157 (1929).

on its own merits, with a view to summarizing certain mathematical methods which are capable of yielding exact information about its solution (though unfortunately they do not suffice for its complete solution). These same methods prove to be of even greater value when combined with the perturbation theory.

To present these methods in as simple a manner as possible, it has seemed desirable to treat them quite independently of the general quantum theory. It is presumed that the reader is familiar with this theory, however. In distinction to the point of view of quantum dynamics, attention will be focused on those characteristic solutions of Eq. (1) for which W has a single definite value. Solutions for different values of W will occasionally be considered, but only after suitable announcement of the fact. The linearity of the operator H has as its consequence that, if $\psi = u$ and $\psi = v$ are solutions of Eq. (1) for the given value of W, so is $\psi = au + bv$ whenever a and b are constants independent of the coordinates.[2] If the boundary conditions are homogeneous, $au + bv$ will also be a characteristic solution if u and v are such; it is assumed that this is the case. These statements may conveniently be summarized by saying that the solutions of Eq. (1) form a linear manifold or space; the last term is justified by the fact that the functions of the linear manifold satisfy the same abstract axioms as the vectors of a space.[3]

In particular, there is an integer ν, called the number of dimensions of the manifold, determined by the statement that any $\nu+1$ members of the manifold are linearly dependent, while it is always possible to find ν members which are linearly independent. Any ν independent members, $u_1 u_2 \cdots u_\nu$, are called a basis for the manifold, and any other member ψ can be expressed as[4]

$$\psi = a_n u_n , \tag{2}$$

where the constants a_n are determinate. The basis corresponds to the set of "unit" vectors along an oblique set of axes in a space, the a_n to the coordinates measured along these axes with the appropriate units. The number ν is also called the degree of degeneracy and the statistical weight of the energy level; the case $\nu = 1$ is of course not excluded, and it is assumed that ν is finite.

It is possible to define a unit of length in this space: it is known that in case the wave equation does not involve the spin coordinates, integrals of the type[5]

[2] An operator R is said to be linear if

$$R(f+g) = Rf + Rg,$$
$$R(af) = a(Rf),$$

where f and g are arbitrary functions on which it is able to act and a is a constant.

[3] H. Weyl; Gruppen theorie und Quantenmechanik, Chap. I (Leipzig, 1928). This book will be referred to as Weyl, *l.c.*

[4] The usual summation convention is adopted.

[5] The asterisk indicates that g^* is the conjugate complex function to g.

$$\int f(x)g^*(x)d\tau, \quad (d\tau = \rho(x)dx_1dx_2 \cdots) \tag{3}$$

play an important part in the theory of its solution; the integral is to be extended over all values of the variables x which have a physical significance. If the spin variables s_i (each of which is capable of assuming only the values $+\frac{1}{2}$ or $-\frac{1}{2}$) do enter into Eq. (1), the role of these integrals is played by

$$\sum_{s_1 s_2} \cdots \int f(x_i s_i)g^*(x_i s_i)d\tau. \tag{3a}$$

To include these two cases it is convenient to represent both by the same notation

$$(f, g). \tag{4}$$

It will be called the scalar product of the functions f and g. The following properties of the scalar product will be verified:

$$\begin{aligned} (f, g) &= (g^*, f^*) = \text{constant}; \\ (af, g) &= a(f, g), \text{ if } a \text{ is a constant}; \\ (f + f', g) &= (f, g) + (f'g); \\ (f, f) &> 0, \text{ unless } f \equiv 0, \text{ in which case it vanishes.} \end{aligned} \tag{5}$$

Because of the last of the properties the scalar product is suitable for a definition of length in the manifold of W: if $(u,u) = c^2$, then $c \geqq 0$ may be called the length of u. Furthermore, if

$$(f, g) = 0$$

then f and g are said to be orthogonal.

One other property of the inner product deserves special mention: if $f = f_1(x_1 \cdots x_n s_1 \cdots s_n) f_2(x_{n+1} \cdots x_n s_{n+1} \cdots s_n)$ and g also has the same form, then

$$(f, g) = (f_1, g_1)_1(f_2, g_2)_2$$

where($)_1$ and ($)_2$ are operations analogous to ( ) but involving only the coordinates $x_1 \cdots x_n s_1 \cdots s_n$ and $x_{n+1} \cdots x_n s_{n+1} \cdots s_n$ respectively. This formula is analogous to breaking up an N-fold integral into the product of an n-fold and $(N\text{-}n)$-fold integral when the integrand can be factored as above.

For the purposes of this review, H may be restricted to the class of hermitian operators, which satisfy the relation

$$(Hf, g) = (f, Hg).$$

This formula indicates that the operator H is such that an operation analogous to an integration by parts can be carried out on the scalar product of Hf and g.

It is evident that the ν functions $u_1 u_2 \cdots u_\nu$ may always be chosen orthogonal to one another, and of unit length. As it will be necessary to use this

fact from time to time, a proof may be given here. Suppose that $u_1 u_2 \cdots u_\mu$, $\mu < \nu$, are orthogonal to one another, and each have the length 1, while v is any other member of the manifold which is linearly independent of $u_1 \cdots u_\mu$. If $u_{\mu+1}$ can then be determined so that it is orthogonal to $u_1 \cdots u_\mu$ and is of unit length, the theorem has, to all intents and purposes, been proven. But this is true, for

$$u' = v - u_1(v, u_1) - u_2(v, u_2) - \cdots - u_\mu(v, u_\mu)$$

is orthogonal to $u_1 \cdots u_\mu$ and cannot vanish, since v is linearly independent of the other functions. If its length is c:

$$(u', u') = c^2,$$

then $u_{\mu+1} = u'/c$ satisfies the required conditions. While this procedure can always be used to construct an orthogonal basis for any given manifold, it is a lengthy computational process, and to be avoided if possible. The basis u_n will therefore not be assumed to be orthogonal unless that fact is explicitly stated.

The manifold of W, hereafter called $\mathfrak{M}_w$, contains other linear manifolds, just as a three dimensional space contains straight lines and planes (one and two dimensional linear manifolds). These may be defined by means of their bases: let $v_1 v_2 \cdots v_\mu$ be $\mu < \nu$ linearly independent functions of $\mathfrak{M}_w$. These generate the submanifold $\mathfrak{M}_1$, whose functions are

$$\psi = \sum_{\alpha=1}^{\mu} a_\alpha v_\alpha$$

and which is μ dimensional.

If one linear submanifold $\mathfrak{M}_1$, of μ dimensions be given, it is always possible to find another, $\mathfrak{M}_2$, of $\mu' \leqq \nu - \mu$ dimensions such that

$$(\psi, \psi') = 0$$

if ψ is in the one and ψ' in the other. The two manifolds are said to be orthogonal. To prove the linearity of $\mathfrak{M}_2$, note that if (ψ, ψ') and (ψ, ψ'') are both zero, $(\psi, a\psi' + b\psi'')$ is also zero. The bases of two orthogonal manifolds must be linearly independent, so that $\mu + \mu' \leqq \nu$. The actual construction may be carried out by a modification of the process of orthogonalization.

If $\mu + \mu' = \nu$, $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are said to be complementary. In this case, any member ψ of $\mathfrak{M}_w$, may be expressed as the sum of two functions ψ_1 and ψ_2 which are contained in $\mathfrak{M}_1$ and $\mathfrak{M}_2$ respectively. They are called the components of ψ in the two complementary submanifolds.

2. **The integrals of the wave equation and the finite matrices representing them.** Quantum dynamics also contemplates linear operators, other than H, which are capable of acting on functions of the coordinates. They are called dynamical variables, but need not necessarily possess analogues among

the variables of classical dynamics. Let R be one of these; if it commutes[6] with H, i.e., if

$$HR = RH \ , \tag{7}$$

it is said to be an integral of the Eq. (1). Integrals possess the following property: if u is a solution of Eq. (1) so is $\psi = Ru$; for

$$H\psi = RHu = HRu = R(Wu) = W\psi .$$

This does not prove that if u is a *characteristic* solution of Eq. (1), Ru will also be characteristic. This is usually the case, however, and may be included in the definition of the term "integral" as a separate requirement. The product of two integrals is obviously again an integral.

Let $R, S, T \cdots$ be integrals of Eq. (1) and $u_1 \cdots u_\nu$ a basis of the manifold of W. Then, because of this property of integrals and the properties of the basis,

$$Ru_n = R_{nm}u_m, \ Su_n = S_{nm}u_m, \text{ etc.}, \tag{8}$$

where the coefficients R_{nm}, S_{nm}, are numbers independent of the coordinates. They may be arranged in matrix array, the left hand index numbering the rows, and the right hand index the columns. Furthermore,

$$RSu_n = S(R_{nm}u_m) = R_{nm}S_{mk}u_k, \tag{9}$$

so that the matrix associated to the product of two operators is itself the product of the matrices associated with the factors of the operator product. *The order of factors in the two products is the same.* Whenever a set of matrices is associated to a set of operators in this manner, they are said to be multiplicatively associated. If, in addition, one and only one matrix is associated to each operator then the set of matrices is called a representation of the set of operators.[7] Both conditions are obviously fulfilled by the matrices (8); the functions u_n are said to be the basis of the representation.

It is not possible to construct a representation of more general dynamical variables using u_n as basis. It is true that it is possible to associate matrices of ν rows and columns to any dynamical variable (by taking the "time average" of the infinite matrix associated to it by quantum theory) but these do not satisfy the requirement that the matrix product be associated to the operator product. The following discussion will therefore apply directly only to integrals.

The definition of the matrices by Eq. (8) is obviously not independent of

[6] If R and S are two operators, their product $T = RS$ is defined by $Tf = g$, where g is determined from

$$Rf = g_1$$
$$Sg_1 = g.$$

The factors of the product act successively in the order of reading from *left to right.* $RS \neq SR$ in general, but it is assumed that $Q(RS) = (QR)S$.

[7] The converse condition, that only one operator be associated to each matrix, need not necessarily be fulfilled. This definition is implied whenever the term representation is used in the following discussion.

the choice of basis. Let $v_1 \cdots v_\nu$ be another basis for the manifold; then, because of the properties of a basis

$$v_\alpha = a_{\alpha n} u_n, \tag{12}$$

and the determinant of the matrix $a_{\alpha n}$ does not vanish. If

$$Rv_\alpha = R'_{\alpha\beta} v_\beta, \tag{13}$$

then

$$R'_{\alpha\beta}\, a_{\beta n} = a_{\alpha m} R_{mn}, \tag{14}$$

as follows at once on substituting Eq. (12) into (13) and using Eq. (8). Since the determinant of $a_{\alpha n}$ does not vanish, its reciprocal exists and Eq. (14) may also be written

$$R'_{\alpha\beta} = a_{\alpha n} R_{nm} a^{-1}{}_{m\beta}. \tag{14a}$$

Representations which are related in this manner are said to be equivalent, and to arise from one another by a canonic transformation with the matrix $a_{\alpha n}$. The accented matrices also constitute a representation of the operators. It is to be emphasized that the matrix $a_{\alpha n}$ is the same for every matrix $R_{mn}, S_{mn}, \cdots$ of the representation. The concept of equivalent representations is fundamental in the following discussion, where equivalent representations are not considered to differ essentially from one another.

3. **Further restriction of the operators to be considered. The group postulates.** In the foregoing paragraph it has been remarked that only integrals of Eq. (1) will be considered in this review; further restrictions are now to be placed on the class of integrals which form the primary objects of study. This will somewhat restrict the generality of the results, but in return for this, the theorems will be much more numerous and specific in their content.

The first restriction is the assumption that the integrals under consideration[8] form a group. Any set of elements is said to be a group[9] if the following three postulates are obeyed:

I. If R and S are members of the group, their products, RS and SR, have a well-defined significance and are also members of the set. The associative law $R(ST) = (RS)T$ is obeyed by the multiplication.

II. An element I, called "the identity" or "unity" and having the properties $RI = IR = R$ for any R, is a member of the group.

III. Each element R uniquely determines another element of the group, called its reciprocal R^{-1} and defined by $R^{-1}R = RR^{-1} = I$.

The first of these postulates is satisfied by the set of all integrals of Eq. (1). The element I may be defined by the relation

[8] It is not assumed that *all* integrals of Eq. (1) satisfy the group postulates, but only that some do. The restriction imposed on the form of Eq. (1) is thus very slight.

[9] H. Weyl, *l.c.* p. 97.

$$If = f; \tag{15}$$

it is a (trivial) integral of Eq. (1).

The third group postulate is more stringent: it excludes all integrals which have no reciprocal and all which have more than one. (The operator $\partial/\partial x$ cannot be a member of any group, since its reciprocal is not unique.)

The theory of abstract groups does not specify the analytic character of the elements of the group: they may be operators, matrices, or ordinary numbers. For the moment, operator groups are under consideration, but it may be remarked that the matrices representing a group of operators often satisfy the group postulates also.

The second restriction to be placed on the integrals is the assumption that they are unitary: i.e., satisfy the relation

$$(R^{-1}f, g) = (f, Rg), \tag{10a}$$

which may also be written

$$(f, g) = (Rf, Rg). \tag{10b}$$

The justification for this name is obtained by noting that f and Rf have the same length: R leaves the unit of length unchanged. The unitary operators are thus analogous to rotations of the vectors of a space. This may be brought out more clearly by considering the matrices representing unitary operators: let u_n be a normallized orthogonal basis for $\mathfrak{M}_w$, so that

$$(u_n, u_m) = \delta_{nm}$$

and

$$Ru_n = R_{nm}u_m,$$
$$R^{-1}u_n = {R^{-1}}_{nm}u_m;$$

then

$$\begin{aligned} {R_{nm}}^{-1} &= (R^{-1}u_n, u_m) \\ &= (u_n, Ru_m), \text{ by (10b)}; \\ &= {R^*}_{mn} \text{ by (5)}. \end{aligned}$$

Hence

$$\delta_{kn} = R_{kn}{R^{-1}}_{nm} = R_{kn}{R_{mn}}^*; \tag{10c}$$

except for the asterisk, these relations are the same as those satisfied by the coefficients of an orthogonal substitution (rotation) of coordinates. A matrix whose elements satisfy Eq. (10c) will be called a unitary matrix. It is to be noted that the matrices representing a set of unitary operators are themselves unitary only when the basis is normallized and orthogonal. According to the definition of the Eq. (14), the representation will always be equivalent to one of unitary matrices; this theorem will be used frequently in the following.

Those integrals of the wave equation which are unitary and form a group will hereafter be called the group $\mathfrak{G}$ of the wave equation, to distinguish them from other integrals and other groups of operators.

4. **Invariant manifolds.** Let ψ be a member of any linear manifold $\mathfrak{M}$ and $I, R, S, T \cdots$ the members of the group $\mathfrak{G}$; if $R\psi$, $S\psi$, $T\psi \cdots$ are also members of $\mathfrak{M}$, it is said to be invariant under the group $\mathfrak{G}$. The manifold $\mathfrak{M}_w$ is invariant under $\mathfrak{G}$; it often happens that a submanifold of $\mathfrak{M}_w$, say $\mathfrak{M}_1$ is also invariant under $\mathfrak{G}$.[10] If this is the case, $\mathfrak{M}_w$ is said to be reducible.

The complementary manifold to $\mathfrak{M}_1$ is also invariant under $\mathfrak{G}$. Let ψ be any member of $\mathfrak{M}_1$, and ψ' of $\mathfrak{M}_2$, the complementary manifold; then by definition

$$(\psi, \psi') = 0.$$

By Eq. (10a), if R is any member of $\mathfrak{G}$,

$$(R^{-1}\psi, \psi') = (\psi, R\psi').$$

But, because of the group postulate III, and the invariance of $\mathfrak{M}_1$, $R^{-1}\psi$ is also in $\mathfrak{M}_1$. Hence the left side of this equation vanishes, and therefore $R\psi'$ is in $\mathfrak{M}_2$, which was to be proven. When the manifolds $\mathfrak{M}_1$ and $\mathfrak{M}_2$ have both been determined, $\mathfrak{M}_w$ is said to have been reduced.

$\mathfrak{M}_1$ and $\mathfrak{M}_2$ may themselves be reducible. If this is the case, the process of reduction may be applied to each, and then to their components in turn, until finally $\mathfrak{M}_w$ has been completely reduced into irreducible submanifolds, $\mathfrak{M}_1, \mathfrak{M}_2, \mathfrak{M}_3, \cdots \mathfrak{M}_q$.

The manifolds $\mathfrak{M}_i$ will each be of a certain number of dimensions, say ν_i, where $\Sigma\nu_i = \nu$. (The possibility $\nu_i = 1$ is, of course, not excluded.) A basis of functions say u_{in}, $(n = 1, 2, \cdots, \nu_i)$ can be found for each; since these functions are independent by definition, and are ν in number when i varies from 1 to q, it follows that they constitute a basis for the whole of $\mathfrak{M}_w$.

The importance of the invariant submanifolds lies in the fact that the whole of the reasoning of Section 2 can be repeated for each of these (in place of $\mathfrak{M}_w$) *in-so-far as only integrals of the group* $\mathfrak{G}$ *are considered.* Each $\mathfrak{M}_i$ is the basis of a representation[11] of the group $\mathfrak{G}$:

$$Ru_{in} = {}^{i}R_{nm}u_{im}; \tag{16}$$

representations which are based on irreducible manifolds are themselves called irreducible.

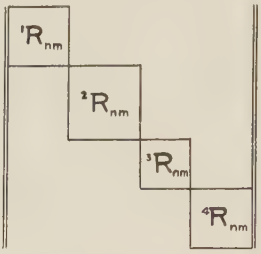
The q matrices (16) may also be considered as a single matrix of ν rows and columns, each row or column being numbered with two indices i n, in the order $11, 12, \cdots, 1\nu_1; 21, 22, \cdots, 2\nu_2; 31 \cdots, q\nu_q$. Analytically, the matrices defined by Eq. (16) have the form

$$R_{in,jm} = {}^{i}R_{nm}\delta_{ij}; \tag{17}$$

[10] Example: Three dimensional Euclidean space is invariant under the group of rotations about the z-axis, the xy plane and the z-axis are invariant submanifolds. They are complementary and irreducible.

[11] Matrices belonging to different representations of the group $\mathfrak{G}$ will be distinguished by an index at the upper left of the letter; *such indices are never summed unless it is explicitly indicated.*

when arranged in matrix array, they have the form:



(17a)

in which all elements not in the squares along the principal diagonal are zero. Every matrix representing an element of the group $\mathfrak{G}$ has this form. As a particular case, all the $\mathfrak{M}_i$ may be one dimensional, in which case every matrix $^iR_{nm}$ is an ordinary number, and (17a) is a diagonal matrix.

Let $w_1 w_2 \cdots w_\nu$ be any other basis of $\mathfrak{M}_w$ and

$$Rw_\alpha = R_{\alpha\beta} w_\beta. \tag{18}$$

The representation $R_{\alpha\beta}$ must be equivalent to the representation (17); this is the significance of the term "reducible" as applied to representations. The set $\mathfrak{M}_w$ is usually given in terms of a basis like w_α and the first problem which presents itself is the calculation of the canonic transformation which transforms it into u_{in}:

$$u_{in} = a_{in,\alpha} w_\alpha. \tag{19}$$

This problem will be solved with considerable generality in the following pages.

It is convenient to introduce another convention into the definition of complete reduction: two or more of the smaller matrices $^iR_{nm}$ may be equivalent. If this is the case, they may always be made identical by a canonic transformation, and this will be assumed to have been accomplished in the process of complete reduction. It is convenient to add a third index to the bases u_{in}, which indicates to which one of a number of identical representations the function belongs. The basis is then designated by u_{ipn}, and

$$Ru_{ipn} = {}^iR_{nm} u_{ipm},$$

where p has any value from 1 to q_i (q_i being the number of times the representation $^iR_{nm}$ appears in (17a)). The Eq. (19) then becomes

$$u_{ipn} = a_{ipn,\alpha} w_\alpha. \tag{19a}$$

5. **Schur's lemmas.** As a preliminary step toward the solution of this problem, certain criteria for the reducibility of a representation will be deduced. The definition of reducibility immediately results in one: let R_{nm} be a reducible representation of $\mathfrak{G}$, whose basis is $u_1 \cdots u_\nu$:

$$Ru_n = R_{nm} u_m. \tag{20}$$

Since the manifold $\mathfrak{M}$ subtended by $u_1 \cdots u_\nu$ is reducible, there exist $\mu < \nu$ functions of the form

$$v_\alpha = a_{\alpha n} u_n, \quad (\alpha = 1, 2 \cdots \mu; n = 1, 2, \cdots \nu;) \tag{21}$$

such that

$$R v_\alpha = R'_{\alpha\beta} v_\beta. \tag{22}$$

On substituting Eqs. (21) and (20) into Eq. (22) and using the independence of the u_n, it follows that

$$R'_{\alpha\beta} a_{\beta m} = a_{\alpha n} R_{nm}. \tag{23}$$

This is the necessary condition that R_{nm} be reducible; it is formally identical with Eq. (14), but it is to be noted that $a_{\alpha n}$ in Eq. (23) is a rectangular matrix of $\mu < \nu$ rows and ν columns, while in Eq. (14), it is a square matrix of ν rows and columns. It is again to be emphasized that this equation is valid for every element R of the group and that $a_{\alpha n}$ is the same matrix for each R.

The Eq. (23) is also the sufficient condition for the reducibility of R_{nm} (if the trivial matrix $a_{\alpha n}$ all of whose elements are zero be excluded) since if a non-trivial matrix exists satisfying Eq. (23), it may always be used to define v_α in Eq. (21), which then satisfy Eq. (22) because of Eq. (20).

Lemma I.

If R_{nm} be an irreducible representation of a group $\mathfrak{G}$ of unitary elements and $R'_{\alpha\beta}$ any other representation of $\mathfrak{G}$ having fewer rows and columns than R_{nm}, then no matrix $a_{\alpha n}$ exists and satisfies Eq. (23) for all R, other than $a_{\alpha n} \equiv 0$.

A second extremely useful theorem is the following:

Lemma II.

If R_{nm} be an irreducible representation of $\mathfrak{G}$ and $R'_{\alpha\beta}$ another representation of $\mathfrak{G}$ having the same number of rows and columns as R_{nm}, while $a_{\alpha n}$ is a square matrix satisfying Eq. (14), then either (1). The determinant of $a_{\alpha n}$ does not vanish and the two representations are equivalent; or (2). Every element of $a_{\alpha n}$ is zero.

The theorem is a matter of definition if the determinant of $a_{\alpha n}$ does not vanish, therefore only the second case need be considered. A matrix $b_{\alpha\beta}$ may be found in any case such that it transforms $a_{\alpha n}$ to diagonal form:

$$b_{\alpha\beta} a_{\beta n} b_{nm}^{-1} = a_\alpha \delta_{\alpha m}. \tag{24}$$

Since the determinant of a matrix is invariant under a canonic transformation, at least one of the a_α must vanish. It is no loss of generality to assume $a_1 = a_2 = \cdots a_\mu = 0$, $a_{\mu+1} \neq 0$, $a_{\mu+2} \neq 0$, $\cdots$, $a_\nu \neq 0$. Substituting in Eq. (14) and rearranging:

$$b_{\gamma\alpha} R'_{\alpha\beta} b_{\beta(k)}^{-1} a_{(k)} = a_{(\gamma)} b_{(\gamma)n} R_{nm} b_{mk}^{-1} \tag{25a}$$

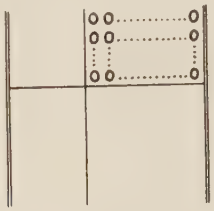
which is not to be summed over γ and k. Calling

$$b_{\gamma n} R_{nm} b_{mk}^{-1} = {}^1R_{\gamma k}, \quad b_{\gamma\alpha} R'_{\alpha\beta} b_{\beta k} = {}^2R_{\gamma k},$$

(these matrices are also representations of the group $\mathfrak{G}$) Eq. (25a) becomes

$$ {}^1R_{\gamma(k)}a_{(k)} = a_{(\gamma)}{}^2R_{(\gamma)k}. \tag{26} $$

This being valid for every R, it follows that ${}^1R_{\gamma k}$ has the form



(27)

since ${}^1R_{\gamma k}=0$ whenever $\gamma \leqq \mu$, $k>\mu$, because of Eq. (26). Now let R_{mn} be based on the functions u_n which, by hypothesis, subtend an irreducible manifold $\mathfrak{M}$. Then the functions

$$ v_\alpha = b_{\alpha n}u_n $$

satisfy the relation

$$ Rv_\gamma = {}^1R_{\gamma\beta}v_\beta $$

and hence, because of (27), $v_1 \cdots v_\mu$ subtend an invariant manifold, which is contrary to hypothesis if $\mu<\nu$. Hence every a_α must be zero, and therefore every $a_{\alpha n}$, which was to be proven.

The properties of irreducible representations were first studied by Frobenius and Burnside.[12] The above proof of the second lemma is based on a more general proof given by I. Schur, and the two theorems may be called Schur's Lemmas, though the theorem proven by Schur is somewhat different and more general than either.[13]

As a corollary to the second lemma, it may be shown that

If R_{mn} is an irreducible representation of $\mathfrak{G}$ and a_{nm} commutes with all R_{mn}:

$$ R_{mn}a_{nk} = a_{mn}R_{nk}, $$

then $a_{nm}=a\delta_{nm}$, where a is independent of n.

A matrix $b_{\alpha\beta}$ exists satisfying Eq. (25), but it is now no longer necessary that one of the a_α vanish. In Eq. (26), ${}^1R_{\gamma k}={}^2R_{\gamma k}$, so that it becomes

$$ (a_k - a_\gamma){}^1R_{\gamma k} = 0 $$

[12] G. Frobenius, Sitzungsberichte der Berliner Akademie, 1896–99. Burnside, Theory of Groups of Finite Order, Camb. (1911).

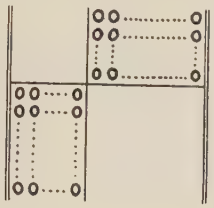
[13] I. Schur, Sitzungsberichte der Berliner Akademie, XVIII (1905). The theorem proven in this paper reads:

Let R_{mn} and $R'_{\alpha\beta}$ be any two irreducible representations of the same group (not necessarily of unitary operators), of ν and ν' rows and columns respectively. If $a_{\alpha n}$ is a matrix satisfying

$$ R'_{\alpha\beta}a_{\beta m}=a_{\alpha n}R_{nm} $$

for all R, then either $a_{\alpha n}=0$, or $\nu=\nu'$ and the determinant of $a_{\alpha n}$ does not vanish.

which is true for all R. Hence if all the a_k are not equal, all ${}^1R_{\gamma k}$ must have the form



which is contrary to the assumption of the theorem; but if all the a_k are equal, then the original matrix must also have the form $a_{nm}=a\delta_{nm}$.

6. **An application to perturbation theory.** The solution of the problem stated at the end of Section 4 may profitably be interrupted here to show the relation of the concept of irreducible manifolds to the problem of solving the wave equation by perturbation methods. Whenever the operator H can be written in the form

$$H = H^0 + \epsilon H',$$

where ϵ is a numerical parameter of given value, both the solutions and characteristic values of the equation

$$(H^0 + \epsilon H')\Psi = W\Psi \tag{28}$$

may be expanded in a power series in ϵ:

$$\begin{aligned} \Psi &= \psi + \epsilon\psi' + \epsilon^2\psi'' + \cdots, \\ W &= W + \epsilon W' + \epsilon^2 W'' + \cdots. \end{aligned} \tag{29}$$

Even when ϵ is not small, the first one or two terms of these series are often fair approximations to their sums.

The successive members of these series satisfy the equations

$$H^0\psi - W\psi = 0, \tag{30}$$

$$H^0\psi' - W\psi' = (W' - H')\psi, \tag{31}$$

$$H^0\psi'' - W\psi'' = (W' - H'\ \psi' + W''\psi, \tag{32}$$

. . . . . .

These may be solved successively for ψ, ψ', ψ'', etc.; each of these functions must itself satisfy the boundary conditions of the problem. The characteristic values and functions of Eq. (30) are assumed known: let them be W_k and $v_{k\alpha}(\alpha=1\cdots\nu)$ respectively. The functions $v_{k\alpha}$ subtend the manifold of the energy level W_k. For the following it will be necessary to adopt the general view-point of quantum theory and consider all possible characteristic values W_k and their functions. The $v_{k\alpha}$ may be supposed to constitute a complete and normalized orthogonal set, so that

$$H'v_{k\alpha} = \sum_l \sum_\beta H'_{k\alpha,l\beta}v_{l\beta}, \tag{33}$$

where

$$H'_{k\alpha,l\beta} = (H'v_{k\alpha}, v_{l\beta}).$$

This matrix contains an infinite number of rows and columns, and possibly the indices k and l may take on all values in a certain continuous range also; this possibility will not be considered.

Now, if

$$\psi = \sum_{\alpha} A_\alpha v_{k\alpha} \tag{34}$$

it will satisfy Eq. (30) for $W = W_k$, and if ψ' be assumed to have the form

$$\psi' = \sum_{l} \sum_{\beta} B_{l\beta} v_{l\beta}, \tag{35}$$

Eq. (31) becomes

$$\sum_{l} \sum_{\beta} B_{l\beta}(W_l - W_k) v_{l\beta} = \sum_{\alpha} A_\alpha W' v_{k\alpha} - \sum_{\alpha} \sum_{l} \sum_{\beta} A_\alpha H'_{k\alpha,l\beta} v_{l\beta}.$$

On equating the coefficients of $v_{l\beta}$ on both sides, it follows that

$$B_{l\beta} = \frac{\sum_{\alpha} A_\alpha H'_{k\alpha,l\beta}}{W_k - W_l} \qquad (k \neq l), \tag{36}$$

and

$$0 = A_\beta W' - \sum_{\alpha} A_\alpha H'_{k\alpha,k\beta}, \qquad (k = l); \tag{37}$$

the last term of Eq. (37) is not to be summed over k.

The Eqs. (36) determine the $B_{l\beta}$ in terms of the A_α, and the Eqs. (37) are ν homogeneous linear equations in the ν unknowns A_α. Unless the determinant of these vanishes, the A_α must all be zero; if

$$\begin{vmatrix} H'_{k1,k1} - W' & H'_{k1,k2} & \cdots & H'_{k1,k\nu} \\ H'_{k2,k1} & H'_{k2,k2} - W' & \cdots & H'_{k2,k\nu} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ H'_{k\nu,k1} & H'_{k\nu,k2} & \cdots & H'_{k\nu,k\nu} - W' \end{vmatrix} = 0 \tag{38}$$

then the A_α need not vanish and their ratios are determinate. Eq. (38) is an algebraic equation of the ν-th degree in W', and is called the secular equation. Its roots are the only values of W' for which Eq. (31) possesses a solution satisfying the boundary conditions of the problem.

The properties of the finite matrices $H'_{k\alpha,l\beta}$ for fixed k and l are now to be determined on the assumption that the elements of the group $\mathfrak{G}$ are integrals of both Eqs. (28) and (30), so that

$$\begin{aligned} RH &= HR, \\ RH' &= H'R, \end{aligned} \tag{39}$$

and

$$Rv_{k\alpha} = {}^{k}R_{\alpha\beta}v_{k\beta},$$

whenever R is a member of $\mathfrak{G}$. From Eq. (33) it follows that

$$H'Rv_{k\alpha} = R\{H'_{k\alpha,l\beta}v_{l\beta}\} = \sum_{l}\sum_{\beta\gamma}H'_{k\alpha,l\beta}{}^{l}R_{\beta\gamma}v_{l\gamma},$$

and

$$RH'v_{k\alpha} = H'\{{}^{k}R_{\alpha\beta}v_{k\beta}\} = \sum_{l}\sum_{\beta\gamma}{}^{k}R_{\alpha\beta}H'_{k\beta,l\gamma}v_{l\gamma}.$$

Comparing with the second of Eqs. (39), it is seen that

$${}^{k}R_{\alpha\beta}H'_{k\beta,l\gamma} = H'_{k\alpha,l\beta}{}^{l}R_{\beta\gamma}, \tag{40}$$

which is not to be summed over k and l. The infinite matrix H' can thus be broken up into finite matrices, each of which satisfies an equation similar to that entering the hypothesis of Schur's lemmas.

If all the representations ${}^{k}R_{\alpha\beta}$ are irreducible, it follows at once that $H'_{k\alpha,l\beta}$ (and therefore $B_{l\beta}$) vanishes unless ${}^{l}R_{\alpha\beta}$ is equivalent to ${}^{k}R_{\alpha\beta}$. Furthermore, the component for which $k=l$ satisfies the condition of the corollary of the last section, and must therefore have the form

$$H'_{k\alpha,k\beta} = H'(k)\delta_{\alpha\beta},$$

where $H'(k)$ is independent of α. The Eqs. (37) and (38) therefore reduce to

$$A'_{\beta}(W' - H'(k)) = 0,$$

and

$$(W' - H'(k))^{\nu} = 0.$$

The ν roots of the secular equation are therefore all equal and the energy level remains ν-fold degenerate, to this approximation at least. By discussing Eq. (32), etc., in the same manner, this result may be proven for each approximation in turn. It is therefore true of the rigorous solution of Eq. (28) also.

If the representations ${}^{k}R_{\alpha\beta}$ are not all irreducible, the situation is more complex. For simplicity, only that portion of the matrix H' for which $k=l$ will be discussed. The subscripts k may then be dropped, so that Eqs. (37) and (40) become

$$0 = A_{\beta}W' - A_{\alpha}\bar{H}'_{\alpha\beta}, \tag{37a}$$

and

$$R_{\alpha\beta}H'_{\beta\gamma} = H'_{\alpha\beta}R_{\beta\gamma}. \tag{40a}$$

The matrix $H'_{\alpha\beta}$ may be referred to as the perturbation matrix; it does not follow from Eq. (40a) that the perturbation matrix is diagonal, but these equations still impose definite restrictions upon its characteristic values. To bring these into evidence the basis v_{kn} may be supposed completely reduced by the transformation

$$u_{ipn} = a_{ipn,\alpha} v_{k\alpha}, \tag{41}$$

so that

$$Ru_{ipn} = {}^{i}R_{nm} u_{ipm}, \tag{42}$$

where ${}^{i}R_{nm}$ is an irreducible representation of $\mathfrak{G}$. The function ψ may now be written

$$\psi = c_{ipn} u_{ipn},$$

where

$$c_{ipn} = A_\alpha a^{-1}_{\alpha,ipn}.$$

Substituting this into Eq. (37a), one obtains

$$c_{jqm} W' - c_{ipn} H'_{ipn,jqm} = 0, \tag{43}$$

where

$$H'_{ipn,jqm} = a_{ipn,\alpha} H'_{\alpha\beta} a^{-1}_{\beta,jqm}.$$

The perturbation matrix thus undergoes a canonic transformation with the matrix a. Eq. (40a) becomes

$${}^{i}R_{nm} H'_{ipm,jqs} = H'_{ipn,jqm} {}^{j}R_{ms}$$

(which is not to be summed over i or j). The perturbation matrix has thus been broken up into smaller matrices (one for each set of values of i, p, j, q) each of which satisfies the hypothesis of one of Schur's lemmas. It therefore follows at once that

$$H'_{ipn,jqm} = H'_{pq} \delta_{ij} \delta_{n,m} \tag{44}$$

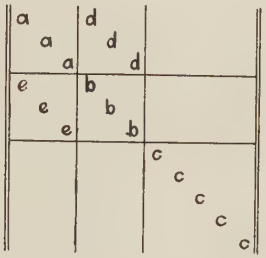
where H'_{pq} is dependent only on i, not on n.

This result may most readily be visuallized by writing the matrices out in tabular form. As an example, let $R_{\alpha\beta}$ have been reduced to the form

	11m	12m	21m
11n	${}^{1}R_{nm}$		
12n		${}^{1}R_{nm}$	
21n			${}^{2}R_{nm}$

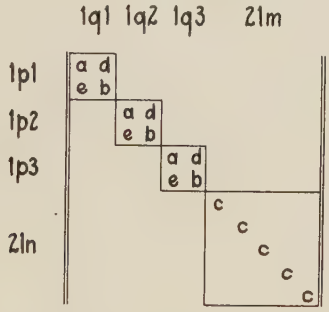
$$= {}^{i}R_{nm} \delta_{ij} \delta_{pq}. \tag{45}$$

The matrices occupying the first two squares along the diagonal are identical, but are inequivalent to the one in the third square. Then the matrix (44) must have the form



$$= H'_{pq}\delta_{ij}\delta_{nm}, \qquad (46)$$

in which all elements not expressly indicated are zero, and all indicated by the same letter are equal. This matrix may be brought into another form by rearranging the order of its rows and columns:



$$= H'_{pq}\delta_{ij}\delta_{nm}. \qquad (47)$$

The special form of the matrix may also be described in terms of the Eqs. (43). These consist of a number of distinct sets, one for each pair of values of j and m. Each set contains just q_j equations and involves only those q_j unknowns c_{jqm} for which j and m have this set of values. Furthermore, the coefficients of those sets having the same j (but different m) are identical. When one set has been solved, the solution of ν_j sets is thus known at once. It is obvious that the characteristic values W' are independent of m so that although the effect of the perturbation has been to remove some of the degeneracy of the unperturbed equation, the energy levels remain degenerate. This degeneracy is such that the functions belonging to a given energy level still form the basis of a representation of the group $\mathfrak{G}$, which will generally be irreducible. When the characteristic functions of a given energy level are the basis of an irreducible representation of $\mathfrak{G}$, the degeneracy is said to be *normal*. When the representation is reducible, the degeneracy is said to be *accidental*. It is seen that any perturbation whose operator commutes with the elements of $\mathfrak{G}$ (so that they are integrals of both the perturbed and unperturbed equation) can remove only the accidental degeneracy, but cannot affect the normal degeneracy. The study of the irreducible representations of $\mathfrak{G}$ is thus equivalent to the study of the normal degeneracy of the energy levels of all equations admitting these operations as integrals.

7. **The multiplication table of a group. The rearrangement theorem.** To use Schur's lemmas effectively in the study of irreducible representations, it is necessary to be able to calculate matrices $a_{\alpha n}$ satisfying Eq. (23) or (14). This is made possible by a theorem which depends on all three of the group postulates of Section 3, and which will be proven in this section.

Thus far, the number of distinct elements (operators) in the group $\mathfrak{G}$ has not entered into consideration. It may contain only a finite number, N, (N is then called the order of $\mathfrak{G}$) or it may contain an infinite number. This infinity may be denumerable or nondenumerable; an infinite group is called discrete or continuous according to these two alternatives. In the applications, a finite group (the permutation group) and an infinite continuous group (the rotations) will be encountered. These two cases will therefore be considered here; infinite discrete groups will not be treated.[14]

For the first, let the group be finite; its elements may then be distinguished by different letters: $I, A\ B\ C \cdots K\ L\ M\ N$ or by indices, $R_1 = I, R_2, R_3, \cdots R_N$.

The first notation will be preferred, since the number of indices to be considered is large as it is. When the elements are arranged in the order $I\ A\ B\ C \cdots M\ N$, they will be said to be in their "natural" order, though this order is as arbitrary as the symbols. It is then possible to construct a multiplication table of N rows and columns; the elements are entered in the first row and first column in their natural order, and the entry at the intersection of the c-th column and r-th row is $R_r R_c$, where R_r is the first element of the r-th row and R_c, the first element of the c-th column. Since $R_r R_c \neq R_c R_r$, this table will usually not be symmetric about the principal diagonal.

Example: The multiplication table of a certain group[15] of six elements $I\ A\ B\ C\ D\ E$ is

I	A	B	C	D	E
A	I	D	E	B	C
B	E	I	D	C	A
C	D	E	I	A	B
D	C	A	B	E	I
E	B	C	A	I	D

From it may be determined the various relations:

$$A^2 = B^2 = C^2 = I,\ BC = D,\ CB = E,\ \text{etc.}$$

This is not the most general multiplication table of six elements, for it will be noted that each element occurs once and only once in each row and each column. This is a direct consequence of the group postulates, and, for our purposes, their most important consequence.

[14] The theory of groups of finite order is treated by
Burnside, *l.c.* A. Speiser, Theorie der Gruppen endlicher Ordnung, (Berlin 1927).
G. A. Miller, H. F. Blichtfeldt, L. E. Dickson, Theory and applications of groups of finite order. (New York, 1916).
The Theory of continuous groups is found in Weyl, Chap. III.

[15] Speiser, *l.c.* p. 12.

It may be proven for the general case of a finite group as follows: suppose the element A does not occur in the column headed by R_c; then there is no element R_r in the first column (and therefore none in the group) for which $R_r R_c = A$. But, by Postulate III, R_r^{-1} is an element of the group and exists, and $X = AR_c^{-1}$ is also a member of the group by Postulate I. Since $XR_c = A$, this contradicts the supposition, which is therefore false. Each of the N elements of the group must thus appear at least once in each column. As there are only N entries in each column, it cannot appear more than once. Similar reasoning proves the theorem for the rows.

This result may be stated in terms of the elements of an invariant manifold. Let u be any member of $\mathfrak{M}$; then Au, Bu, etc. are also members, as is

$$v = \sum_R a(R)Ru. \tag{48}$$

In this formula, the $a(R)$ are ordinary constants, the notation indicating that a different number is associated to each element R of $\mathfrak{G}$; $a(R)$ may be called a numerical function of the elements of $\mathfrak{G}$, or, more simply, a function of R.[16] The summation indicates that R is to be given, successively, the significance of I, A, B, $C \cdots M$, N and the resulting N functions $a(R)Ru$, added together to form the function v. It is also possible to write

$$T = \sum_R a(R)R \tag{49}$$

where the operator T is defined by

$$v = Tu;$$

this defines the addition of operators in the usual manner. T is not always a member of $\mathfrak{G}$, however, though it is an integral of Eq. (1).

The importance of the theorem regarding the multiplication table is that it enables the expression

$$Sv = \sum_R a(R)RSu \tag{50}$$

to be rearranged into

$$Sv = \sum_R a(RS^{-1})Ru. \tag{51}$$

The significance of this equation may best be made clear in the case of an example based on the group I, A, B, C, D, E. Eq. (48) then reads

$$v = a(I)u + a(A)Au + a(B)Bu + a(C)Cu + a(D)Du + a(E)Eu.$$

If $S = D$, Eq. (50) becomes

$$Dv = a(I)Du + a(A)Bu + a(B)Cu + a(C)Au + a(D)Eu + a(E)Iu$$

[16] This terminology may conflict with that current in quantum theory, according to which a function of an operator or set of operators is itself an operator; where confusion is to be avoided, $a(R)$ will be called a numerical function of R. The most general operator function of the elements of a group is a linear one.

upon using the fifth column of the multiplication table for the expressions RD. The order of the terms may be rearranged into

$$Dv = a(E)u + a(C)Au + a(A)Bu + a(B)Cu + a(I)Du + a(D)Eu.$$

This is of the same form as the original expression except that the coefficient of, say Cu, is not $a(C)$ but $a(B)$, where $B = CE = CD^{-1}$. The action of the operator D on v results in a permutation of the coefficients $a(R)$. This is true only if D is a member of the group $\mathfrak{G}$, and only because each operator appears just once in each row and column of the multiplication table. If the first were not true, it would not be possible to represent Dv as a linear function of u, Au, $Bu \cdots$; if the first were true and the second not, then the coefficients of this linear function would not be so simply related to those of the original function v.

The equality of the sums (50) and (51) may be called the rearrangement theorem.[17] It may be extended without difficulty to the case of discrete infinite groups; the sum (48) is then an infinite series, and the theorem is certainly true if the series (50) converges absolutely. The extension to continuous groups requires some preliminary discussion of their multiplication table. The elements of a continuous group cannot be distinguished by integral indices, but must be considered as determined by continuously variable parameters. Let these be $\alpha \equiv (\alpha_1\alpha_2 \cdots)$; then the elements of the group may be represented by $R(\alpha_1\alpha_2 \cdots) \equiv R(\alpha)$. The significance of this notation is that

$$R(\alpha)u(x) = v(\alpha, x).$$

The result of operating on a function $u(x)$ is a function[18] of both α and x, which will be assumed to be sufficiently regular at all points of a certain range of α to make possible the analytic operations required in the following. For example, let the operations of $\mathfrak{G}$ be rotations of a plane geometrical figure about an axis perpendicular to its plane, and (x, y) the rectangular coordinates of a point in the figure. Then the parameter α may conveniently be taken to be the angle of the rotation. If $u = y^2$, then

$$R(\alpha)y^2 = (x \sin \alpha + y \cos \alpha)^2.$$

The association of the numerical values of the parameters to the elements of the group is to a large extent arbitrary (α, in the example, may be measured either in degrees or radians; $\beta = \alpha^2$ might also have been chosen as parameter). In order to be useful, however, this assignment must satisfy the condition of uniqueness: to every set of values $(\alpha_1\alpha_2 \cdots)$ in a certain range Γ there must correspond one and only one element of the group, and no element of the group may correspond to more than one set of values of α

[17] This theorem is usually proven in a more abstract form in textbooks on group theory and is not referred to by this name. See Speiser, p. 15.

[18] It may be noted that if R_1, $R_2 \cdots$ are elements of a finite group,

$$R_n u(x) = v(n, x),$$

so that this equation introduces nothing fundamentally new.

in Γ. In order that the angle α of the example may satisfy this requirement, the range Γ must be defined by $0 \leqq \alpha < 2\pi$ (if α is measured in radians). Γ is called the range of the group.

The multiplication table of the continuous group may now be given an analytic expression, even though a tabular array is impossible: if

$$R(\alpha'') = R(\alpha)R(\alpha'), \tag{52}$$

then the multiplication table is known as soon as α'' is known as a function of α and α':

$$\begin{aligned} \alpha_1'' &= \phi_1(\alpha_1\alpha_2 \cdots ; \alpha_1'\alpha_2' \cdots) \\ \alpha_2'' &= \phi_2(\alpha_1\alpha_2 \cdots ; \alpha_1'\alpha_2' \cdots) \\ &\cdots \cdots \cdots \cdots \end{aligned} \tag{53}$$

In the case of the example,

$$\alpha'' = \alpha + \alpha', \text{ mod } 2\pi.$$

The theorem regarding the multiplication table of the finite group may now be extended to continuous groups:

The equations $\alpha'' = \phi(\alpha, \alpha')$ possess the following properties:

1. If α' (or α) be held constant and α (or α') allowed to vary over the whole range Γ, then α'' also varies over the whole of Γ, nor does it assume the same value twice if α (or α') does not.

2. The equations may be solved both in the sense $\alpha = \phi'(\alpha', \alpha'')$, and $\alpha' = \phi''(\alpha, \alpha'')$, and the functions ϕ' and ϕ'' possess the same properties as ϕ.

The first of these statements may be proven by reasoning identical with that used for the finite group, the second is really implied in the first but also follows directly from the uniqueness of the reciprocal.

The analogue to Eq. (48) is now

$$v = \int_\Gamma a(\alpha)R(\alpha)u(x)d\Omega \text{ where } d\Omega = \sigma(\alpha)d\alpha_1 d\alpha_2 \tag{54}$$

The function $\sigma(\alpha)$ may be left arbitrary for the moment, except that it be single valued and finite in Γ. If $S = R(\alpha')$, then Eq. (50) becomes

$$R(\alpha')v = \int_\Gamma a(\alpha)R(\alpha)R(\alpha')u(x)d\Omega \tag{55}$$

or, on changing form α to α'' as the variables of integration,

$$R(\alpha')v = \int_\Gamma a(\alpha)R(\alpha'')u(x)\frac{\sigma(\alpha)}{\sigma(\alpha'')}\frac{\partial(\alpha_1\alpha_2 \cdots)}{\partial(\alpha_1''\alpha_2'' \cdots)}d\Omega'',$$

in which α is supposed to be expressed as a function of α' and α'' by means of proposition (2) above. *This integral is again to be extended once over the range*

Γ because of the properties of ϕ and ϕ'; this would not be the case if propositions (1) and (2) were not true. The expression may be simplified if $\sigma(\alpha)$ be determined so that

$$\frac{\sigma(\alpha)}{\sigma(\alpha'')} \frac{\partial(\alpha_1\alpha_2 \cdots)}{\partial(\alpha_1''\alpha_2'' \cdots)} = 1; \tag{56}$$

it then reads

$$R(\alpha')v = \int_\Gamma a(\alpha)R(\alpha'')u(x)d\Omega''. \tag{57}$$

The identity of the Eqs. (55) and (57) is the rearrangement theorem for continuous groups.

In the following, it will be convenient not to make any formal distinction between finite and continuous groups. This is possible if it be agreed that the notation of Eq. (48) shall also represent the integral of Eq. (54), with $a(R) \equiv a(\alpha)$, in the case of continuous groups.

The rearrangement theorem may obviously be extended to the matrices of a representation of $\mathfrak{G}$: if[19]

$$T_{nm} = \sum_R a(R)R_{nm},$$

then

$$\begin{aligned} T_{nm}S_{mk} &= \sum_R a(R)R_{nm}S_{mk} \\ &= \sum_R a(RS^{-1})R_{nk}. \end{aligned}$$

8. **The orthogonality relations.** Let ${}^1R_{nm}, {}^2R_{nm}, \cdots, {}^iR_{nm}, \cdots$, be irreducible representations, no two of which are equivalent; each is based on a closed manifold $\mathfrak{M}_i$ of ν_i dimensions. If $b_{\beta m}$ be any matrix of ν_i rows and ν_j columns, the matrix

$$a_{\alpha n} = \sum_R {}^iR^{-1}_{\alpha\beta}b_{\beta m}{}^iR_{mn}$$

satisfies the relation

$${}^iS_{\alpha\beta}a_{\beta m} = a_{\alpha n}{}^iS_{nm}$$

for all operators S in the group. This may be proven as follows:

$$a_{\alpha n}{}^iS_{nm} = \sum_R \sum_k c_{\alpha k}(R){}^iR_{kn}{}^iS_{nm},$$

where

$$c_{\alpha k}(R) = {}^iR^{-1}_{\alpha\beta}b_{\beta k}.$$

On applying the rearrangement theorem to the summation over R:

[19] Schur, *l.c.*, calls a matrix defined in this manner a "group matrix."

$$a_{\alpha n}{}^{j}S_{nm} = \sum_{R} \sum_{k} c_{\alpha k}(RS^{-1})\,{}^{i}R_{km}.$$

Now

$$c_{\alpha k}(RS^{-1}) = {}^{i}S_{\alpha\beta}c_{\beta k}(R),$$

and hence

$$a_{\alpha n}{}^{j}S_{nm} = {}^{i}S_{\alpha\beta}\sum_{R} \sum_{k} c_{\beta k}(R)\,{}^{i}R_{km}$$

$$= {}^{i}S_{\alpha\beta}a_{\beta m},$$

which was to be proven. Hence by Schur's lemmas, $a_{\alpha n}$ must vanish whenever $i \neq j$ and have the form $a(i)\delta_{\alpha n}$ when $i = j$. Since the matrix $b_{\beta m}$ is entirely arbitrary, this is equivalent to

$$\sum_{R} {}^{i}R^{-1}_{\alpha\beta}\,{}^{j}R_{mn} = \gamma_{\beta m}\delta_{ij}\delta_{\alpha n}, \tag{58}$$

where the matrix $\gamma_{\beta m}$ is independent of α and n. This matrix may readily be determined: setting $i = j$, $\alpha = n$ in Eq. (58) and summing over n from 1 to ν_i, the result is

$$\sum_{R} {}^{i}R^{-1}_{n\beta}R_{mn} = \sum_{R} \delta_{m\beta} = \nu_i\gamma_{\beta m}.$$

The second sum is equal to

$$N\delta_{m\beta}$$

where N is the order of $\mathfrak{G}$ if it is a finite group, or

$$N = \int_{\Gamma} d\Omega$$

if $\mathfrak{G}$ is continuous. Hence Eq. (58) becomes [20]

$$\sum_{R} {}^{i}R^{-1}_{\alpha\beta}\,{}^{j}R_{mn} = \frac{N}{\nu_i}\delta_{ij}\delta_{\alpha n}\delta_{\beta m}. \tag{59}$$

If these representations have been chosen so that their matrices are all unitary, $({}^{i}R^{-1}_{\alpha\beta} = {}^{i}R^{*}_{\beta\alpha})$ then Eq. (59) may also be written

$$\sum_{R} {}^{i}R^{*}_{\beta\alpha}\,{}^{j}R_{mn} = \frac{N}{\nu_i}\delta_{ij}\delta_{\alpha n}\delta_{\beta m}. \tag{59a}$$

In the case of a continuous group, each element ${}^{i}R_{\alpha\beta}$ (for given i, α, β) is a function of the group parameters only; Eq. (59a) states that these functions are all orthogonal to one another. In the case of a finite group, the elements ${}^{i}R_{\alpha\beta}$ (for fixed i, α, β again) are vectors in a certain space of N dimensions (the "group manifold"), and Eq. (59a) states that these vectors are orthogonal.

[20] The arrangement of the indices on the two sides of this equation is to be noted carefully. Reference: Weyl, *l.c.* Chap III.

For this reason the Eq. (59a), and the more general Eq. (59) as well, will be referred to as the orthogonality relations between the representations. They constitute a powerful analytic tool for the solution of complex problems.

An important corollary may be deduced from these equations in the case of finite groups: since there can be at most N orthogonal vectors in a space of N dimensions, and the ${}^iR_{\alpha\beta}$ constitute $\sum_{all\ i}\nu_i^2$ such vectors, it follows that this sum must be less than or at most equal to N. Since $\nu_i^2 \geqq 1$, it follows that *there are at most a finite number of inequivalent irreducible representations of a finite group*. The problem of finding *all* inequivalent irreducible representations of a finite group thus acquires a definite significance, and admits of solution. In the case of continuous groups, the orthogonality relations also materially restrict the irreducible representations, though their number is infinite. All inequivalent irreducible representations of the three dimensional rotation group will be determined in the special part of this review.

The converse of the theorem is also true: if a given sequence of representations ${}^iR_{\alpha\beta}$ satisfies Eq. (59) then they are irreducible and inequivalent. To prove this, let $A_{\alpha m}$ be a rectangular matrix satisfying

$$ {}^iR_{\alpha\beta}A_{\beta m} = A_{\alpha n}{}^jR_{nm} $$

for all R. On multiplying both sides by ${}^iR^{-1}_{fq}$ and summing over all R, Eq. (59) results in

$$ A_{pm}\delta_{q\alpha} = A_{\alpha q}\delta_{pm}\delta_{ij}\,. $$

If $i \neq j$, A_{pm} must be zero because the right side vanishes; if $i=j$, $A_{pm}=A\,\delta_{pm}$, where A is independent of p or m. Hence, the representation ${}^iR_{nm}$ is irreducible.

9. **The complete reduction of a given invariant manifold.** Let the sequence of representations ${}^1R_{nm}$, ${}^2R_{nm}\cdots$ of the previous section be known to contain all inequivalent irreducible representations of $\mathfrak{G}$. Then the task of completely reducing a given invariant manifold $\mathfrak{M}$ can be accomplished by a series of multiplications and summations (or integrations in the case of a continuous group). Let $\mathfrak{M}$ be given in terms of its basis $w_1, w_2, \cdots w_\nu$, so that

$$ Rw_\alpha = R_{\alpha\beta}w_\beta\,. \tag{60} $$

Then, if $a_{m\alpha}$ be any arbitrary matrix of ν_i rows and ν columns, the ν_i functions

$$ v_{in} = \sum_R {}^iR^{-1}_{nm}a_{m\alpha}R_{\alpha\beta}w_\beta \tag{61} $$

satisfy the relations

$$ Sv_{in} = {}^iS_{nm}v_{im} \tag{62} $$

if S is any member of $\mathfrak{G}$. To prove this, it is merely necessary to substitute from Eq. (61) on the left side of Eq. (62) and apply the rearrangement theorem.

Hence, *if the functions v_{in} do not vanish identically* they subtend an irreducible invariant submanifold of $\mathfrak{M}$. Repeating the process for various values of i and different a_{mn}, all irreducible submanifolds of $\mathfrak{M}$ can be found. Furthermore, that basis in each which is adapted to the matrices ${}^iR_{nm}$ of the standard representation is automatically obtained. This process smacks somewhat of trial and error, however, for often the functions v_{in} will vanish identically, and other times the same invariant submanifold will be obtained. A still more complicated situation arises when several nonvanishing bases for the same representation are obtained: let v_{in}, v'_{in}, v''_{in} all be bases for ${}^iR_{mn}$. Then it is quite possible that

$$v_{in} = av'_{in} + bv''_{in}$$

where a and b are independent of n. There is no general way of excluding such [21] linear relations, and therefore this method of reduction is not always the most practicable. Less laborious methods can often be found in special cases.

One general result may be deduced, however, for the completely reduced basis u_{ipn} which is determined in this manner may be written

$$u_{ipn} = \sum_R {}^iR^{-1}_{nm} a^p_{m\alpha} R_{\alpha\beta} w_\beta, \tag{61a}$$

where the matrices $a^p_{m\alpha}$ are arbitrary except for the restriction that the u_{ipn} be ν in number and linearly independent. The coefficients $a_{ipn,\alpha}$ of Eq. (19a) are therefore given by

$$a_{ipn,\alpha} = \sum_R {}^iR^{-1}_{nm} a^p_{m\beta} R_{\beta\alpha} .$$

10. **The character of a representation.** The element of trial and error in the method of reduction just outlined will be materially reduced if some general method can be found for determining those representations ${}^iR_{nm}$ which appear in the completely reduced form of a given representation, and the number of times q_i which each is repeated. (cf. Eq. (17a). This problem is solved by the theory of characters.[22]

This theory is based on the proposition that the sum of the diagonal elements or "spur" of a matrix is equal to the spur of any matrix arising from it by a canonic transformation: i.e., that

$$\sum_{n=1}^{\nu} R_{nn} = \sum_{\alpha=1}^{\nu}{}' R_{\alpha\alpha} . \tag{63}$$

[21] Linear relations of the form

$$av_{in} + bv_{jm} = 0$$

cannot exist when $i \neq j$, $m \neq n$. For on operating with R,

$$a^iR_{nk}v_{ik} + b^jR_{ml}v_{jl} = 0;$$

on multiplying this by ${}^iR^{-1}_{nk}$ and summing over all R, Eq. (59) results in $a = 0$.

[22] Weyl, l.c. p. 117.

whenever

$$'R_{\alpha\beta} = a_{\alpha n} R_{nm} a^{-1}_{m\beta}. \tag{64}$$

This proposition is obvious on setting $\alpha = \beta$ in Eq. (64) and summing, for

$$\sum_{\alpha=1}^{\nu} a_{\alpha n} a^{-1}_{m\alpha} = \delta_{nm}.$$

Hence, the numerical function of the operators R which is defined by

$$\chi(R) = \sum_{n=1}^{\nu} R_{nn} \tag{65}$$

is the same for all representations $'R_{\alpha\beta}$ which are equivalent to R_{nm}. The converse of this proposition is also true for irreducible representations, as will be shown. The function $\chi(R)$ is therefore called the character or characteristic of the representation R_{nm}. If R_{nm} is irreducible, its character is said to be primitive.

On setting $\alpha = \beta$, $m = n$ in Eq. (59) and summing over α and m, the following important property of primitive characters becomes evident:

$$\sum_R {}^i\chi(R^{-1})\,{}^j\chi(R) = N\delta_{ij}. \tag{66}$$

(The indices i indicate the representation of which ${}^i\chi(R)$ is the character.) Since every representation of the group $\mathfrak{G}$ is equivalent to a unitary one, for which ${}^iR^{-1}_{nm} = {}^iR^{*}_{mn}$, it follows that

$${}^i\chi(R^{-1}) = [{}^i\chi(R)]^*$$

so that Eq. (66) may be written

$$\sum_R {}^i\chi^*(R)\,{}^j\chi(R) = N\delta_{ij}. \tag{68}$$

This is called the orthogonality relation between primitive characters. The converse of the above proposition follows at once from this equation; for suppose $i \neq j$, but ${}^i\chi(R) = {}^j\chi(R) = \chi(R)$ for all R. Then

$$\sum_R |\,\chi(R)\,|^2 = 0$$

by Eq. (68), whence $\chi(R) = 0$ for all R, including $R = I$. But $\chi(I) = \sum_n I_{nn} = \nu$, and hence the number of rows and columns in the matrices of the two representations would have to be zero—an absurd result.

The problem proposed in the first prargraph of this section can now be solved with considerable elegance: let $R_{\alpha\beta}$ be equivalent to a completely reduced representation of the form (17a), in which ${}^iR_{mn}$ appears q_i times. (If a particular irreducible representation j does not appear, $q_j = 0$). Then

$$\chi(R) = \sum_\alpha R_{\alpha\alpha} = \sum_i q_i\,{}^i\chi(R);$$

on multiplying this equation by ${}^{i}\chi^{*}(R)$ and summing over R, Eq. (68) yields

$$\sum_{R} {}^{i}\chi^{*}(R)\chi(R) = Nq_{i}. \tag{69}$$

The computational work involved in calculating the character of $R_{\alpha\beta}$ and the sums (69) is usually less than that involved in determining by trial which of the sums (61) yields a nonvanishing set of v_{in}. Furthermore the knowledge of the number of independent sets of v_{in} to be expected is a definite advantage, and is sometimes sufficient to solve a given problem without actually carrying out the reduction of the manifold $\mathfrak{M}$.

The function $\chi(R)$ possesses one general property which materially reduces the amount of work required for its calculation; the numerical values of $\chi(R)$ and $\chi(S)$ are equal when R and S belong to the same "class" of the group $\mathfrak{G}$. A "class" of the elements belonging to a given group is defined as follows: if R and T are in the group, so is $S = TRT^{-1}$, and S is said to belong to the same class as R, or to be equivalent to R. It may also be said that S arises from R by means of a canonic transformation with a third element T of the group. Since the matrix S_{nm} is equal to $T_{nk}R_{kl}T^{-1}_{lm}$, the equation $\chi(R) = \chi(S)$ follows from the proposition regarding the invariance of the diagonal sum under a canonic transformation.

The number of classes, N_c, in a given finite group cannot be greater than the order of the group, and there are thus only $N_c \leqq N$ different numbers required to specify each character. As all primitive characters are linearly independent because of Eq. (69), it follows that there are at most N_c such characters, and hence at most N_c inequivalent irreducible representations. This restricts the number of irreducible representations of a finite group still more than Eq. (59).

11. **Miscellaneous theorems.** I. Any basis u_n of a unitary irreducible representation of $\mathfrak{G}$ is orthogonal, except for a normalizing factor. For

$$\begin{aligned}(u_n, u_m) &= (Ru_n, Ru_m)\\ &= R_{nk}R_{ml}{}^{*}(u_k, u_l);\end{aligned}$$

summing over R and using Eq. (59a) which applies to unitary irreducible representations,

$$N(u_n, u_m) = \delta_{nm} \sum_{k} (u_k, u_k)$$

or

$$(u_n, u_m) = C\delta_{nm},$$

where C is independent of n.

II. If two groups are given, which have no common element other than I, and if $QR = RQ$ whenever R belongs to one group and Q to the other, they can be combined into a single group whose elements are (1) all the elements R of one group; (2) all the elements Q of the other; (3) all products of the

form RQ. The resulting group is called the direct product of the two given groups.

If R_{nm} and $Q_{\alpha\beta}$ are irreducible representations of the two groups, the matrices $R_{nm}Q_{\alpha\beta}$ (whose rows are numbered with the two indices n,α and columns with m,β) form an irreducible representation of their direct product. This is most readily proven by noting that the elements of these matrices satisfy the orthogonality relations (59).

III. If a function u is given, which can be subjected to all the operations of a group $\mathfrak{G}$, it is possible to construct a linear manifold of functions invariant under $\mathfrak{G}$. For let R, S, be any operations of $\mathfrak{G}$, then the manifold $\mathfrak{M}$ may be defined by the statement that Ru and Su are always members, as is any linear function of its members. The most general function of this manifold is representable as

$$\psi = \sum_R a(R)Ru.$$

The linearly independent among the functions Ru may be chosen as its basis. It is said to be generated by u and $\mathfrak{G}$.

If all the functions Ru are linearly independent, it follows that v_{in} (Eq. (61)) cannot vanish for any i or any $a_{m\alpha}$. There will therefore be precisely ν_i different bases for the representation ${}^iR_{nm}$ contained in $\mathfrak{M}$.

IV. If the elements of two groups can be multiplicatively associated (cf. Section 2) in such a manner that every element of either group corresponds to one and only one element of the other, the groups are said to be isomorphic. Their multiplication tables are identical if one abstracts from superficial differences in notation and arrangement. It is even possible to abstract from more fundamental differences: the elements of one group may be operators, those of the other matrices. The matrices representing an operator group sometimes (but not always, cf. reference 7) form an isomorphic group. It is obvious that any representation of a group is also a representation of any group isomorphic to it. This theorem will be exceedingly useful in the following.

V. If R_n is a ν-dimensional irreducible representation and

$$\chi_{nm} = \sum_{class} R_{nm},$$

where the summation is not to be extended over the whole group, but only over a single class of elements, then

$$\chi_{nm} = \frac{N_c}{\nu}\chi(R)\delta_{nm},$$

R being any of the members of this class, and N_c the number of members in the class if the group is finite; if the group is continuous,

$$N_c = \int_{class} d\Omega.$$

Proof: if S is any element of the group

$$S_{nm}\chi_{mk}S^{-1}_{kl} = \sum_{class} S_{nm}R_{mk}S^{-1}_{kl}.$$

But $SRS^{-1} = T$ belongs to the same class as R.
Hence

$$S_{nm}\chi_{mk}S^{-1}_{kl} = \sum_{class} T_{nl} = \chi_{nl}.$$

The matrix χ_{nm} thus commutes with all members of the representation, and must therefore have the form $a\delta_{nm}$ where a is independent of n. The constant a may be evaluated by taking the diagonal sum:

$$va = \sum_{class}\chi(R) = N_c\chi(R).$$

This theorem has been used by Dirac[23] as a definition of the characters $\chi(R)$.

II. Specification of the Groups to be Considered

12. **The integrals in the absence of an external field. The case of one electron.** In Section 3, the class of integrals under consideration was restricted by certain formal postulates. In this section, integrals are to be determined which satisfy these conditions. For the first, the spin variables s may be omitted from consideration,[24] and only the translatory variables considered; then if the system consists of single electron moving in a centra field of force, the wave equation must be invariant under a rotation of coordinate axes about an axis through the center of force. This fact is sufficient to determine some of the integrals of the group $\mathfrak{G}$.

This may be proven in a somewhat general form: let $H(x)$ and $f(x)$ be an arbitrary operator and function, respectively, involving the variables $x = x_1x_2\cdots$, and

$$x_i = \phi_i(x'_i) \tag{70}$$

a substitution of variables. Let R denote the operation of performing this substitution on any function of x; then

$$\begin{aligned} R[H(x)f(x)] &= H(\phi(x'))f(\phi(x')) \\ &= H'(x')f'(x'). \end{aligned} \tag{71}$$

If $H'(x) = H(x)$, the operator H is said to be invariant under the substitution R. Suppose this to be the case, and that $f(x)$ satisfies the equation

$$H(x)f(x) = Wf(x); \tag{72}$$

then

$$R[H(x)f(x)] = H(x')f'(x') = Wf'(x').$$

[23] P. A. M. Dirac, Proc. Roy. Soc. **A123,** 721 (1929).
[24] E. Wigner, Zeits. f. Physik **43,** 624; **45,** 601 (1927).

If the accent on the variables x be dropped after the substitution it is seen that $f'(x)$also satisfies Eq. (72) for the same value of W as $f(x)$. With a slight modification of the definition of the operation R, $f'(x)$ may be written $Rf(x)$, and R is then an integral of the Eq. (72).

The operation R may also be given a geometrical interpretation, for Eq. (70) relates every point x' (Fig. 1) to another point $x=\phi(x')$ of the same space

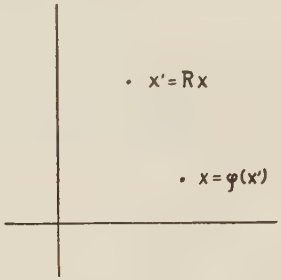


Fig. 1.

in such a manner that when either is given, the other may be determined. The function $f'(x)=Rf(x)$ assumes the same numerical value at $x=x'$ as $f(x)$ assumes at $x=\phi(x')$. The point x' may also be denoted by Rx, whereupon $f'(Rx)=f(x)$.

In the present case, the functions (70) are linear and homogeneous in the coordinates $x'y'z'$:

$$\begin{aligned} x &= a_1x' + a_2y' + a_3z' \\ y &= b_1x' + b_2y' + b_3z' \\ z &= c_1x' + c_2y' + c_3z' \end{aligned} \tag{73}$$

which represents a rotation of coordinate axes if

$$\begin{aligned} a_1b_1 + a_2b_2 + a_3b_3 &= 0, \quad \text{etc.}, \\ a_1^2 + a_2^2 + a_3^3 &= 1, \quad \text{etc.}, \\ \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} &= +1. \end{aligned} \tag{74}$$

The coefficients a_n, b_n, c_n, may all be expresseed in terms of the three Eulerian angles α, β, γ of the rotation (cf. Section 19). The operation R is to be defined as before and will be denoted by $R(\alpha, \beta, \gamma)$.

As a simple example, consider the Schrödinger equation for a free electron. It possesses the solution $f=\exp ikz$, where $h^2k^2/8\pi^2m=W$. Since it is invariant under the substitution (73), it therefore also possesses the solution $\exp ik(c_1x+c_2y+c_3z)$.

It is geometrically evident that the rotations form a group. They are also unitary operators when the inner product is defined as in Eq. (3) or (3a), as may be verified by performing the substitution of variables (73) in the integrals.

Before considering the general case of the wave equation involving all four coordinates x, y, z, s, it will be well to examine functions of s alone. The most general function of a variable whose only possible values are $+1/2$ and $-1/2$ may always be written in the form

$$\psi = A\xi(s) + B\eta(s) \tag{75}$$

where A and B are independent of s, and

$$\begin{aligned} \xi(\tfrac{1}{2}) &= 1, \quad \xi(-\tfrac{1}{2}) = 0, \\ \eta(\tfrac{1}{2}) &= 0, \quad \eta(-\tfrac{1}{2}) = 1. \end{aligned} \tag{76}$$

The most general linear operation, Q, which may be performed on such a function is a linear substitution of the quantities ξ and η. If a, b, c, d are arbitrary constants and

$$\begin{aligned} \xi &= a\xi' + b\eta' \\ \eta &= c\xi' + d\eta', \end{aligned} \tag{77}$$

then Q is defined by

$$Q\psi = A(a\xi + b\eta) + B(c\xi + d\eta). \tag{78}$$

If ψ is also a function of x, y, z, then it is merely necessary to consider A and B as functions of these variables. Such functions can be acted on by any operator T which is defined so as to be able to act on functions of x, y, z, and also by Q. The most general operation which may be peformed on them is a sum of products like TQ. It is to be noted that $TQ = QT$.

The wave equation of a spinning electron is then

$$H\psi = W\psi,$$

where H is an operator of this general form. Its specific expression need not concern us: its only property which is important is that it possesses integrals of the form $R(\alpha, \beta, \gamma)D(\alpha, \beta, \gamma)$, where $D(\alpha, \beta, \gamma)$ is a particular operation like Q. The form of D may be deduced directly from the postulates of the Pauli or Dirac spin theories.[25] The result may be stated without proof:

$$D: \qquad \begin{aligned} \xi &= a\xi' + b\eta' \\ \eta &= -b^*\xi' + a^*\eta' \end{aligned} \tag{79}$$

where

$$aa^* + bb^* = 1$$

and

$$\begin{aligned} a &= \cos \beta/2 \, e^{i(\alpha+\gamma)/2} \\ b &= \sin \beta/2 \, e^{i(-\alpha+\gamma)/2}. \end{aligned} \tag{80}$$

[25] P. A. M. Dirac, Proc. Roy. Soc. **A117**, 610; **118**, 351 (1928).
W. Pauli, Zeits. f. Phys. **43**, 601 (1927).
P. Jordan, Zeits. f. Phys. **44**, 21 (1927).
J. V. Neumann and E. Wigner, Zeits. f. Phys. **47**, 203 (1928), **49**, 73 (1928).
F. Möglich, Zeits. f. Phys. **48**, 852 (1928).

The quantities a and b have been used by Cayley and Klein[26] in the study of the rotation of a rigid body, and are often called the Cayley-Klein parameters.

The operations D (though not integrals) also satisfy the group postulates and are unitary operators. The last remark may again be verified by performing the substitution in Eq. (13a). To prove that they form a group, it may be noted that the expressions (80) for a and b are the most general consistent with $aa^* + bb^* = 1$, provided only that α, β, γ are entirely arbitrary. Furthermore, if D_{st}, D'_{tr} be two substitutions of the form (79), their resultant substitution $D''_{sr} = D_{st}D'_{tr}$ is again of the form (79), with

$$\begin{aligned} a'' &= aa' - bb'^* \\ b'' &= ab' + ba'^*. \end{aligned} \tag{81}$$

These equations may also be written in terms of α, β, γ, etc. If there were no relation between the a and b of the operations D, and the α, β, γ of R, the operators $R(\alpha\beta\gamma)D(ab)$ would form the direct product (cf. Section 11, theorem II) of the groups R and D. However, the operators here considered do not include all the products $R(\alpha, \beta, \gamma)D(ab)$, but only those for which a and b are expressed in terms of α, β, γ by Eq. (80). It is not obvious from what has been said here that this subset forms a group, but this fact follows from the considerations of Neumann, l. c. An independent proof will be given in Section 19.

These are not the only members of the group $\mathfrak{G}$ in the absence of an external field. If the spin be again neglected for a moment, the wave equation must be invariant when the coordinate system is changed from right- to left-handed. This transformation may most readily be accomplished by the substitution

$$K: \qquad \begin{aligned} x &= -x', \\ y &= -y', \\ z &= -z'. \end{aligned} \tag{82}$$

Let K denote the operation of performing this substitution; then $K^2 = \mathrm{I}$ and $KR = RK$, as is evident from the geometry involved. The operation K is then also a member of the group $\mathfrak{G}$ of the spin-free equation. It remains to be considered how K is to be extended to apply to the spin-variable s. It might be expected that it would be accompanied by a substitution like (78), but this is not the case. This may be made plausible by noting that the variables s are analogous to an angular momentum:

$$xp_y - yp_x = m(x\dot{y} - y\dot{x}).$$

Since this is an even function of the coordinates, the substitution (82) will not change it. The angular momentum is a rotor, not a vector. By analogy, then, it is defined that

$$Kf(x, y, z, s) = f(-x, -y, -z, s), \tag{83}$$

and this operator is also an integral of the Eq. (77).

[26] F. Klein and A. Sommerfeld, Theorie des Kreisels.

The complete group $\mathfrak{G}$ for the wave equation of a single spinning electron in a central field of force thus consists of the direct product of the groups RD and I, K.

13. **Extension to the case of N electrons.** The extension of the foregoing results to the case of N electrons moving under their mutual actions and that of a central field, is readily accomplished since all N electrons are identical and act on each other with forces dependent only on their relative positions and velocities. The operation R is to be defined as the substitution (73) applied to all N coordinate sets x_i, y_i, z_i simultaneously; the same applies to K. The operation D is also readily generalized, for the most general function involving N spin-variables s_i can always be expressed as a sum of 2^N terms, of which

$$\psi(x_i, y_i, z_i)\xi(s_1)\eta(s_2)\eta(s_3)\cdots\xi(s_N) \tag{84}$$

is typical. Each pair of functions $\xi(s_i)$, $\eta(s_i)$ will then undergo the substitution (79). That the operations RD as thus generalized are integrals of the Pauli wave equation will again be assumed without proof. The Dirac equation for one electron has not yet been generalized to the case of N electrons.

These operators obviously form a group which is isomorphic to that considered in the last section and thus possesses the same representations as that group (cf. Section 11, Theorem IV). It does not constitute the whole of the group of integrals in this case, however.

The identity of the N electrons introduces further integrals into the group $\mathfrak{G}$, for the wave equation must be invariant when the coordinates of two electrons are interchanged. If $\psi(x_1, y_1, z_1, s_1; x_2, y_2, z_2, s_2; \cdots,)$ is a solution, $\psi(x_2, y_2, z_2, s_2; x_1, y_1, z_1, s_1; \cdots,)$ will also be a solution. More generally, any permutation of the coordinates of the electrons will result in another solution. Denote these operations by P; they form a finite group of order $N!$. As the permutations commute with the operations of the groups already enumerated, the whole group $\mathfrak{G}$ is simply their direct product.

14. **The integrals in the presence of an external field.** When the system of electrons is acted on by an homogeneous field in addition to the central field of the nucleus and their mutual actions, the equation will no longer be invariant under all rotations of the coordinate system, but only under those rotations whose axes coincide in direction with the external field. This direction may be taken as the z-axis, and $S(\omega)$ may denote a rotation about this axis through the angle ω. The permutations P will still be integrals, since the external field acts on each electron according to the same law.

The operation K is no longer an integral either, but another operation may be found to replace it in the group $\mathfrak{G}$. This operation is different for electric or magnetic fields. In the latter case, the wave-equation is invariant under the transformation

$$K': \begin{aligned} x_i &= x_i', \\ y_i &= y_i', \\ z_i &= -z_i', \end{aligned} \tag{85}$$

which represents a reflection of the coordinate system in the xy-plane. It will be evident that $K'^2 = I$ and $K'S(\omega) = S(\omega)K'$. The group $\mathfrak{G}$ then consists of the direct products of the groups $PS(\omega)$ and I, K'.

In the presence of an electric field along the z-axis, neither K nor K' are integrals, but the operation

$$K'': \begin{aligned} x_i &= x_i', \\ y_i &= -y_i', \\ z_i &= z_i', \end{aligned} \tag{86}$$

which represents a reflection in the xz-plane, is an integral. It differs from K' in that it does not commute with $S(\omega)$. To see this, let I be the point whose coordinates are (x, y), K'' the point whose coordinates are $(K''x, K''y)$, S the point (Sx, Sy), etc. These points are plotted in Fig. 2. The

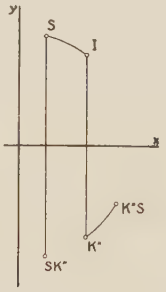


Fig. 2.

points SK'' and $K''S$ do not coincide, but it is seen that $SK'' = K''S^{-1}$. This relation is obviously general. As $K''^2 = I$, it follows that all products (e.g., $S(\omega)K''S(\omega_1)$) can be expressed either in the form $K''S(\omega)$ or else as $S(\omega)$. In the example, $S(\omega)K''S(\omega_1) = K''S^{-1}(\omega)S(\omega_1) = K''S(\omega_1 - \omega)$. Therefore the operations, K'', $S(\omega)$, $K''S(\omega)$ form a group, even though K'' and $S(\omega)$ do not commute. The group $\mathfrak{G}$ is the direct product of this and the permutation group. When the spin of the electrons is taken into account, it is fairly obvious that the rotation $S(\omega) \equiv R(\omega, 0, 0)$ (cf. Section 19) is to be replaced by $S(\omega)D(\omega, 0, 0)$, where $D(\omega, 0, 0)$ is the unitary substitution associated to $R(\omega, 0, 0)$.

If the external field is not homogeneous, or consists of crossed electric and magnetic fields, no rotations are integrals, nor is there any operation analogous to K. The group $\mathfrak{G}$ then consists only of the permutations.

III. Construction of Irreducible Representations of the Rotation Group, etc.

15. **The group I, K.** The irreducible representations of the finite group of order $N = 2$ which contains the identity I and[27] the element K, and whose multiplication table is $K^2 = I$, can be found without difficulty. They are

[27] The element K may be replaced by K' or K'' without altering the validity of these considerations.

two in number (for there are two classes in the group) and because of Eq. (59b) each has but one row and column in its matrices. They are: $I=1$, $K=1$ and $I=1$, $K=-1$. The reduction of any given manifold $\mathfrak{M}$ by the method of Section 9 is simple: let u be any member of $\mathfrak{M}$, then the sums

$$\sum_R {}^i R_{nm}^{-1} R u$$

are easily evaluated:

$$v = u + Ku, \quad w = u - Ku,$$

are functions satisfying

$$Kv = v, \quad Kw = -w,$$

respectively. The tests for linear independence will always be simple and there is little more to be said about this group. It is an almost trivial example of the application of the foregoing methods.

16. **The two-dimensional rotations.** The next simplest of the groups we have discussed in Part II is the group of rotations about the z-axis. Its operators are defined by

$$S\psi(xyz) = \psi(x'y'z'), \tag{87}$$

where

$$\begin{aligned} x &= x' \cos\omega + y' \sin\omega, \\ y &= -x' \sin\omega + y' \cos\omega, \\ z &= z'. \end{aligned}$$

The members of this group all commute with one another:

$$S(\omega)S(\omega_1) = S(\omega_1)S(\omega) = S(\omega_1 + \omega),$$

the sum $\omega+\omega_1$ being taken mod 2π. Any group whose members have this property is called Abelian, and its irreducible representations are necessarily ordinary numbers. For suppose $S_{nm}(\omega)$ were an irreducible representation whose matrices had more than one row and column, then

$$S_{nm}(\omega)S_{mk}(\omega_1) = S_{nm}(\omega_1)S_{mk}(\omega),$$

and therefore

$$S_{nm}(\omega) = a(\omega)\delta_{nm}$$

by the corollary to Schur's lemmas. This is in contradiction to the assumption of irreducibility.

Let $a(\omega)$ represent the operator $S(\omega)$. Then

$$a(\omega)a(\omega_1) = a(\omega + \omega_1),$$

which is a functional equation having as its solution

$$a(\omega) = e^{im\omega}. \tag{88}$$

Because $S(\omega+2\pi)=S(\omega)$, the constant m can take on only integral values (positive or negative). (Cf. the condition of unique correspondence, Section 2.)

The integral over the group is

$$\frac{1}{2\pi}\int_0^{2\pi}\cdots d\omega$$

and the orthogonality relations are

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-im\alpha}e^{in\alpha} = \delta_{nm}.$$

The characters are in this case identical with the "matrices" $a(\omega)$.

17. **The representation of $\mathfrak{G}$ in the case of an external field.** The irreducible representations of the groups which dominate the theory of external fields are readily constructed. The case of a magnetic field may be treated first. The group $\mathfrak{G}$ then consists of the direct product of the groups I, K, and $S(\omega)$. Its representations are therefore two in number (cf. Section 11, theorem II) and can be written down immediately as

$$S(\omega)\rightarrow e^{im\omega}, \quad K'\rightarrow +1,$$

and

$$S(\omega)\rightarrow e^{im\omega}, \quad K'\rightarrow -1. \tag{89}$$

The group $\mathfrak{G}$ in the case of an homogeneous electric field is not Abelian, as has been seen in Section 14. Its representations will therefore not all be one-dimensional. To find them, let $\mathfrak{M}$ be a given manifold which is invariant under $\mathfrak{G}$, and suppose its basis given in such a form that any member of it, say u_m, satisfies the relation

$$S(\omega)u_m = e^{im\omega}u_m.$$

It is always possible to find such a basis, since $\mathfrak{M}$ is also invariant under the subgroup $S(\omega)$. The function $v_m = K'' u_m$ will then satisfy the relation

$$S(\omega)v_m = e^{-im\omega}v_n,$$

which follows at once from $K''S=S^{-1}K''$. The submanifold subtended by the two functions u_m, v_m is invariant and, in general, irreducible. The representation based on it is two-dimensional:

$$S(\omega)\rightarrow \begin{Vmatrix} e^{im\omega} & 0 \\ 0 & e^{im\omega} \end{Vmatrix}, \quad K''\rightarrow \begin{Vmatrix} 0 & 1 \\ 1 & 0 \end{Vmatrix}. \tag{90}$$

In the particular case $m=0$, this submanifold is reducible and (90) breaks up into two one-dimensional representations:

$$S(\omega)\rightarrow 1, \quad K''\rightarrow +1,$$

and

$$S(\omega)\rightarrow 1, \quad K''\rightarrow -1. \tag{91}$$

18. **The unitary substitutions of two variables: $D(\alpha\beta\gamma)$.** It would be logical to consider the three-dimensional rotation group at this point, but it is expedient to consider the group of unitary substitutions of two variables $D(\alpha\beta\gamma)$ first. These two groups are intimately related, as will be shown in the next section, and the analysis of $D(\alpha\beta\gamma)$ is simpler.[28]

The operations $D(\alpha\beta\gamma)$ may be considered as defined by Eqs. (78), (79) and (80). For the present purpose, the special significance of ξ and η is immaterial and they may be taken to be any two complex variables which undergo this substitution. It is unitary, and hence

$$\xi\xi^* + \eta\eta^* = \xi'\xi'^* + \eta'\eta'^*, \tag{92}$$

as may be verified by performing the substitution (91) on the left hand side of (92). The reciprocal of the matrix (79) is therefore

$$\begin{Vmatrix} a^* & -b \\ b^* & a \end{Vmatrix}. \tag{93}$$

The substitution may be performed on any function of ξ and η, not only on linear ones:

$$Df(\xi\eta) = f'(\xi\eta), \tag{94}$$

where

$$f'(\xi'\eta') = f(a\xi' + b\eta', -b^*\xi' + a^*\eta').$$

The multiplication table of the group has already been discussed (cf. Eq. (81)); as a particular case, it will be verified that

$$D(\alpha\beta\gamma) = D(\gamma 00)D(0\beta 0)D(\alpha 00). \tag{95}$$

The group is obviously a three parameter continuous one, and it is necessary to determine the range Γ of its parameters. An inspection of Eq. (80) will convince the reader that if

$$0 \leqq \alpha < 4\pi, \qquad 0 \leqq \beta < \pi, \qquad 0 \leqq \gamma < 2\pi \tag{96}$$

the quantities a and b will assume all pairs of values satisfying $aa^* + bb^* = 1$, and assume each set of values only once. The inequalities (96) thus determine Γ (cf. Section 7).

The classes of elements $D(\alpha\beta\gamma)$ are readily determined. It is a well-known theorem, that if a unitary matrix D_{st} be given, another unitary matrix of determinant $+1$ can be found such that when D_{st} is subjected to a canonic transformation with the latter, it is transformed into a diagonal matrix. As the second matrix is a member of the group here under consideration, it follows that D_{st} belongs to the same class as a matrix

$$\begin{Vmatrix} a' & 0 \\ 0 & a'^* \end{Vmatrix}. \tag{97}$$

[28] Weyl, *l.c.* p. 107.

Since this matrix must be unitary, it follows that $a' = \exp i\omega/2$; because of the invariance of the spur of a matrix

$$a + a^* = a' + a'^* = 2\cos\omega/2,$$

or because of Eq. (80)

$$\cos\omega/2 = \cos(\beta/2)\cos(\alpha+\gamma)/2. \tag{98}$$

The class to which $D(\alpha\beta\gamma)$ belongs is thus determined uniquely by ω, and also contains $D(\omega, 0, 0)$. This result will be of value when the characters of the irreducible representations are to be calculated.

To obtain the irreducible representations, it is necessary to find invariant linear manifolds of functions of ξ and η. The simplest of these are the homogeneous polynomials of νth degree in ξ, η. They constitute a $\nu+1$ dimensional linear manifold $\mathfrak{M}_\nu$ which is subtended by the basis

$$u_p = \xi^p\eta^q \qquad (p+q=\nu,\ p=0, 1, 2\cdots\nu). \tag{99}$$

The manifold $\mathfrak{M}_\nu$ is invariant under the substitutions D, for

$$Du_p = (a\xi + b\eta)^p(-b^*\xi + a^*\eta)^q \tag{100}$$

is again a homogeneous polynomial of νth degree. In particular,

$$D(\omega, 0, 0)u_p = e^{i(p-q)\omega/2}u_p = e^{i(2p-\nu)\omega/2}u_p. \tag{101}$$

The manifold $\mathfrak{M}_\nu$ is irreducible; for suppose it contains an invariant submanifold $\mathfrak{M}_f$ and that

$$f = \sum_{p=0}^{\nu} a_p u_p$$

is a member of this submanifold. Then

$$D(\omega, 0, 0)f = \sum_{p=0}^{\nu} a_p e^{i(2p-\nu)\omega/2}u_p$$

is also a member, no matter what value ω may be given; hence the functions $a_p u_p$ are individually members of $\mathfrak{M}_f$. The submanifold $\mathfrak{M}_f$ contains every u_p for which $a_p \neq 0$ in the polynomial f. Therefore there is certainly one value of p for which

$$D(0, \pi/2, 0)u_p = \left(\frac{\xi+\eta}{\sqrt{2}}\right)^p\left(\frac{-\xi+\eta}{\sqrt{2}}\right)^q$$

is in $\mathfrak{M}_f$; this is a polynomial in which the coefficient of $u_\nu = \xi^\nu$ is certainly not zero. Hence $\mathfrak{M}_f$ contains u_ν and therefore also

$$D(0, \pi/2, 0)u_\nu = \left(\frac{\xi+\eta}{\sqrt{2}}\right)^\nu,$$

which is a polynomial of νth degree *none of whose coefficients vanish.* Hence $\mathfrak{M}_f$ contains every one of the $\nu+1$ functions u_p and therefore coincides with

$\mathfrak{M}_\nu$. It is thus possible to find an irreducible representation for every integral value of $\nu \geqq 0$; it may be shown that there are no others.

The notation of Eq. (99) is not the most convenient for the following; let $\nu = 2l$, $p = l+m$, $q = l-m$, and let the functions

$$u^l_m = \xi^{l+m}\eta^{l-m}/[(l+m)!(l-m)!]^{1/2} \tag{102}$$

be chosen as a basis of $\mathfrak{M}_\nu$. Since ν and p take on integral values, it follows that

$$\begin{aligned} &l = 0, \tfrac{1}{2}, 1, 3/2, \cdots, \\ &m = l - p, \text{ where } p \text{ is an integer} \leqq 2l. \end{aligned} \tag{103}$$

The division brought about by the integral and half-integral values of l and m will prove of importance later. The quantity l will be used as an index for the representation:

$$D(\alpha\beta\gamma)u^l_m = {}^lD_{mn}(\alpha\beta\gamma)u^l_n \tag{104}$$

where the numbers ${}^lD_{mn}$ are defined by the algebraic equation

$$\frac{(a\xi + b\eta)^{l+m}(-b^*\xi + a^*\eta)^{l-m}}{[(l+m)!(l-m)!]^{1/2}} = \sum_{n=-l}^{l} {}^lD_{mn}\frac{\xi^{l+n}\eta^{l-n}}{[(l+n)!(l-n)!]^{1/2}}. \tag{105}$$

The reader will readily verify, for example, that the matrix ${}^1D_{mn}$ has the form

$$\left\| \begin{matrix} a^2 & \sqrt{2}ab & b^2 \\ -\sqrt{2}ab^* & (aa^* - bb^*) & \sqrt{2}a^*b \\ b^{*2} & -\sqrt{2}a^*b^* & a^{*2} \end{matrix} \right\|. \tag{106}$$

The matrix of Eq. (79), which was the starting point of the discussion, is properly denoted by ${}^{1/2}D_{mn}(\alpha\beta\gamma)$.

The matrices ${}^lD_{mn}$ are unitary. To prove this, it must be shown that

$$\sum_{m=-l}^{l} |\, u^l_m \,|^2$$

is invariant; but this is true, for the expression is identically equal to

$$(\xi\xi^* + \eta\eta^*)^{2l}/(2l)!;$$

hence

$${}^lD^{-1}_{mn} = {}^lD_{nm}{}^*. \tag{107}$$

The elements of these matrices, considered as functions of α, β, γ, have numerous properties which are readily deduced from their definition.[29] The most fundamental of these is (cf. Eq. (52) for notation)

$${}^lD_{mn}(\alpha''\beta''\gamma'') = {}^lD_{mk}(\alpha'\beta'\gamma'){}^lD_{kn}(\alpha\beta\gamma) \tag{108}$$

which expresses the fact that they represent the group D. Since the representation is irreducible, it follows from Eqs. (59) and (96) that

[29] They may also be obtained as solutions of a partial differential equation.

$$\int_{\alpha=0}^{4\pi}\int_{\beta=0}^{\pi}\int_{\gamma=0}^{2\pi} {}^{l}D_{mn}{}^{\lambda}D^{*}_{\mu\nu}\sigma d\alpha d\beta d\gamma = \frac{N}{2l+1}\delta_{l\lambda}\delta_{m\mu}\delta_{n\nu}. \tag{109}$$

which is a most important orthogonality relation. The function σ can be determined from the considerations of Section 7, and is found to be

$$\sigma = (\sin\beta)/(4\pi)^2. \tag{110}$$

When σ is normalized with the factor $1/(4\pi)^2$, the value of N is unity. The characters are also readily obtained:

$${}^{l}\chi(\omega) = \sum_{m=-l}^{l} {}^{l}D_{mm}(\omega, 0, 0) = \sum_{m=-l}^{l} e^{im\omega} = [\sin(2l+1)\omega/2]/[\sin\omega/2]. \tag{111}$$

They satisfy the orthogonality relations:

$$\int_{\alpha=0}^{4\pi}\int_{\beta=0}^{\pi}\int_{\gamma=0}^{2\pi} {}^{l}\chi(\omega)^{\lambda}\chi(\omega)\sigma d\alpha d\beta d\gamma = \delta_{l\lambda}, \tag{112}$$

in which ω is to be expressed in terms of $\alpha\beta\gamma$ by means of Eq. (91). Two of the integrations of this triple integral may be performed, leaving

$$\int_{0}^{4\pi} {}^{l}\chi(\omega)^{\lambda}\chi(\omega)\rho(\omega)d\omega = \delta_{l\lambda} \tag{113}$$

where $\rho(\omega)$ is to be determined by integrating σ over all values of $\alpha\beta\gamma$ for which ω has a given value. The reader may verify that

$$\rho(\omega) = (\sin^2\omega/2)/4\pi. \tag{114}$$

On substituting Eq. (80) into Eq. (105), it is seen that

$${}^{l}D_{mn}(\alpha\beta\gamma) = e^{im\gamma}P^{l}{}_{mn}(\beta)e^{in\alpha}, \tag{115}$$

in which P^{l}_{mn} is independent of α and γ. These functions are generalizations of the spherical harmonics. This may be shown by setting $m=0$, $\xi=\eta=1$ in Eq. (105); it then reduces to

$$[\cos\beta + i\sin\beta\sin\alpha]^{l}/l! = \sum_{n} P^{l}{}_{0n}e^{in\alpha}/[(l+n)!(l-n)!]^{1/2}. \tag{116}$$

On multiplying by $\exp(-in\alpha)$ and integrating over α from zero to 2π, it is seen that

$$\frac{l!}{[(l+n)!(l-n)!]^{1/2}}P^{l}{}_{0n}(\beta) = \frac{1}{2\pi}\int_{0}^{2\pi}[\cos\beta + i\sin\beta\sin\alpha]^{l}e^{-in\alpha}d\alpha, \tag{117}$$

which is Laplace's integral for the tesseral harmonics.

19. **The three dimensional rotation group.** The rotation group may be defined analytically by the linear substitutions of Eqs. (73) and (74), the coefficients of which may be expressed in terms of the three Eulerian angles, $\alpha\beta\gamma$. These have a simple geometrical significance: consider a

sphere whose center is at the origin, and on which a quadrant AB is marked By the original orientation of the sphere is meant that in which AB occupies the position————(Fig. 3); when the sphere is rotated, AB will be moved to some position– – –. Let a point P be rigidly attached to the sphere, so that its coordinates in the original orientation are x, y, z, in the rotated orientation, x', y', z'. Then these two sets of coordinates will be related by an equation like (73).

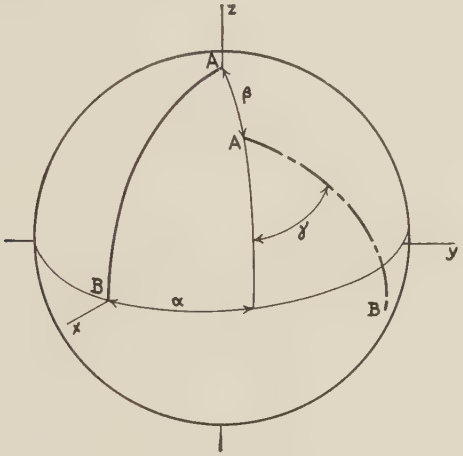


Fig. 3

The most general rotation of the sphere may be resolved into a succession of rotations about the y and z axes (which are fixed in space). Denote a rotation about the z axis through an angle α by $S(\alpha)$, a rotation about the y-axis through an angle β by $T(\beta)$. The angles are to be reckoned positive in the counter-clockwise direction when viewed from a point on the positive axis in question. The substitutions corresponding to these rotations are

$$S(\alpha): \quad \begin{aligned} x &= x' \cos\alpha + y' \sin\alpha \\ y &= -x' \sin\alpha + y' \cos\alpha \\ z &= z'; \end{aligned} \tag{118}$$

$$T(\beta): \quad \begin{aligned} x &= x' \cos\beta - z' \sin\beta \\ y &= y' \\ z &= x' \sin\beta + z' \cos\beta. \end{aligned} \tag{119}$$

Then it is obvious from the figure that the most general rotation is obtained if the sphere be rotated through $S(\gamma)$, then through $T(\beta)$, and finally through $S(\alpha)$:

$$R(\alpha\beta\gamma) = S(\gamma)T(\beta)S(\alpha). \tag{120}$$

The three angles thus specified are the Eulerian angles of the rotation. If they be restricted by the inequalities

$$0 \leqq \alpha < 2\pi, \quad 0 \leqq \beta < \pi, \quad 0 \leqq \gamma < 2\pi, \tag{121}$$

they are uniquely determined by the rotation, and conversely. The Eqs. (120) and (121) are very analogous to Eqs. (95) and (96) for the group of unitary substitutions, the only fundamental difference being that α in (96) is restricted only to lie between 0 and 4π, while in (121), it must lie between 0 and 2π.

The relationship between the two groups[30] which was mentioned at the beginning of Section 18 can now be made clear in a very simple manner: consider the three real functions of the two complex variables ξ, η

$$\begin{aligned} x &= \xi^*\eta + \xi\eta^* \\ y &= \frac{1}{i}(\xi^*\eta - \xi\eta^*) \\ z &= \xi\xi^* - \eta\eta^*. \end{aligned} \tag{122}$$

The identity $x^2+y^2+z^2=(\xi\xi^*+\eta\eta^*)^2$ will readily be verified. Now let ξ, η be subjected to the substitution $D(\alpha\beta\gamma)$ of Eq. (91), and x', y', z' calculated according to Eq. (122) from ξ' and η'. Then x', y', z' are related to x, y, z by a certain linear transformation which is readily calculated in terms of a, b. It will prove to be precisely $R(\alpha\beta\gamma)$. This is most readily seen by noting that $D(\alpha, 0, 0)$ induces the substitution $S(\alpha)$, $D(0, \beta, 0)$ the substitution $T(\beta)$. Furthermore, if two unitary transformations, D and D', induce two substitutions R and R' of the variables xyz, then the product DD' will induce RR'. The result then follows at once from Eqs. (95) and (120). From this it follows that a multiplicative correspondence exists between the elements $R(\alpha\beta\gamma)$ of the rotation group and the elements $D(\alpha\beta\gamma)$ of the unitary substitutions.[31] The correspondence is not unique, however, for if a rotation is induced by $D(\alpha, \beta, \gamma)$ it is also induced by $D(\alpha+2\pi, \beta, \gamma)$. This is the significance of the difference between (96) and (121). The irreducible representations of the rotation group can now be picked out of the set of representations of the group D, for all the matrices ${}^lD_{mn}(\alpha\beta\gamma)$ satisfy the condition of multiplicative correspondence with $R(\alpha\beta\gamma)$, and since (by Eq. (115))

$${}^lD_{mn}(\alpha + 2\pi, \beta, \gamma) = e^{+2\pi i n}\, {}^lD_{mn}(\alpha, \beta, \gamma),$$

all representations whose index l is integral satisfy the condition of unique correspondence. These are therefore the representations of the rotation group, those for which l is a half-integer are not.

The operations $R(\alpha, \beta, \gamma)$ $D(\alpha, \beta, \gamma)$ are seen to form a group which is isomorphic to $D(\alpha, \beta, \gamma)$. They therefore admit all matrices ${}^lD_{mn}(\alpha, \beta, \gamma)$ as representations.

IV. Applications to the Wave Equation

To complete the foregoing chapter, it would be logical to treat the permutation group as completely as the rotation group, etc. This would be a

[30] Weyl, *l.c.* p. 108.

[31] From this and Eq. (98) it may be deduced that every rotation belongs to the same class as a certain rotation $S(\omega)$ about the z-axis, a result which may also be proven geometrically.

lengthy and complex discussion, and is not necessary for present purposes. The brief discussion of the permutation group which is necessary will be given in Chapter V. The present chapter is to be devoted to applications which are independent of the permutation group.

20. **The wave equation without spin terms.** Consider an atom in a field-free region. It may contain any number N of electrons, but their spin is not to be taken into account. Let the rectangular coordinates of the i-th electron be x_i, y_i, z_i, as before. Then the group $\mathfrak{G}$ of the wave-equation contains a sub-group which is composed of the rotations $R(\alpha\beta\gamma)$, the reflection K, and the products $KR=RK$. Let u_m^l be a basis for an irreducible manifold of wave-functions belonging to the energy W, then

$$\begin{aligned} Ru_m^l &= {}^lD_{mn}u_n^l, \quad l = 0, 1, 2, \cdots \\ Ku_m^l &= \pm u_m^l. \end{aligned} \tag{123}$$

The basis will be called even or odd according as the positive or negative sign applies in the last equation.

The problem is to determine as much as possible regarding $u_m{}^l$ from Eq. (123). If the atom contained only one electron, it is known that its solutions have the form[32]

$$u_m^l = X(r)P_l{}^m(\theta)e^{-im\phi} \tag{124}$$

where r, θ, ϕ are the polar coordinates of the electron. Since the ${}^lD_{mn}$ are generalizations of the spherical harmonics, the presumption is that there exists a generalization of this formula which applies to any number of electrons.[33] To discover it, we transform from the variables $x_1y_1z_1 \cdots x_N y_N z_N$ to a new set which consists of three angles, A, B, Γ, specifying the orientation of the atom and $3(N-1)$ other variables specifying the configuration of the electrons relative to each other. This transformation can be accomplished in the following manner: suppose the electrons to be at the points $x_1y_1z_1 \cdots x_Ny_Nz_N$ and imagine them rigidly fastened to one another, so that the whole system can be rotated as a rigid body about any axis passing through the origin. Let it be so rotated until the electron 1 lies on the positive z-axis, and the electron 2 in the first or second quadrant of the xz-plane. Call this rotation $R^{-1}(\mathrm{AB}\Gamma)$; the coordinates of the electrons after the rotation will be

$$0, 0, \bar{z}_1; \bar{x}_2, 0, \bar{z}_2; \bar{x}_3, \bar{y}_3, \bar{z}_3; \cdots \bar{x}_N, \bar{y}_N, \bar{z}_N; (\bar{z}_1 \geqq 0, \bar{x}_2 \geqq 0). \tag{125}$$

These $3(N-1)$ quantities determine the relative positions, and together with A, B, Γ, are single valued functions of the x_i, y_i, z_i. They can therefore be used as coordinates in place of the latter, and will be denoted by A, B, Γ, $\bar{x}$.

In terms of these coordinates, certain functions satisfying the first of Eqs. (123) can easily be written down. They are*

[32] E. C. Kemble, Phys. Rev. Suppl. **1**, 195 (1929), also ftn. on p. 215, *ibid.*

[33] E. Wigner. Zeits. f. Physik **43**, 640 (1927).

* The Eq. (126) may also be deduced rigorously from Eq. (61a).

$$u_m^l = \sum_{n=-l}^{l} {}^lD_{mn}^{-1}(\mathrm{A}, \mathrm{B}, \Gamma) f_n(\bar{x}) = \sum_{n=l}^{l} e^{-i(m\mathrm{A}+n\Gamma)} P_{mn}^l(B) f_n(\bar{x}), \tag{126}$$

where the $2l+1$ functions f_n are any arbitrary functions dependent only on $\bar{x}$; this is the required generalization of Eq. (124). To prove this statement, note that a rotation $R(\alpha\beta\gamma)$ does not alter the relative coordinates, but changes $\mathrm{AB}\Gamma$ into $\mathrm{A'B'}\Gamma'$, where

$$R(\alpha, \beta, \gamma) R^{-1}(\mathrm{A'}, \mathrm{B'}, \Gamma') = R^{-1}(\mathrm{A}, \mathrm{B}, \Gamma). \tag{127}$$

Then if

$$R(\alpha, \beta, \gamma,) u_m^l = v_m^l,$$

$$v_m^l(\mathrm{A'}, \mathrm{B'}, \Gamma', \bar{x}) = u_m^l(\mathrm{A}, \mathrm{B}, \Gamma, \bar{x}) = {}^lD_{mn}(\alpha, \beta, \gamma) u_n^l(\mathrm{A'}, \mathrm{B'}, \Gamma', \bar{x}),$$

by Eqs. (127) and (126).

In this, it has been tacitly assumed that $N \geqq 2$; the particular case $N = 1$ requires special consideration; here the special rotation $R^{-1}(0, \theta, \phi)$ (where θ and ϕ are the polar coordinates of Eq. (124)) brings the electron into the position 0, 0, r. It is then sufficient to set

$$u_m^l = {}^lD_{m0}^{-1}(0, \theta, \phi) f(r) = e^{-im\phi} P^l{}_{m0}(\theta) f(r);$$

the reasoning proving that u_m^l satisfies Eq. (123) is then the same as before.[34]

It remains to consider the second of Eqs. (123). Were it not for the requirements that $\bar{z}_1 \geqq 0$, $\bar{x}_2 \geqq 0$ the function Ku_m^l could be obtained from u_m^l by changing the signs of the variables $(\bar{x})$ and leaving A, B, Γ unaltered. These requirements are essential, however, since otherwise A, B, Γ, $\bar{x}$ would not be uniquely determined. Under these circumstances, it is necessary to resolve K into a rotation through π radians about the $\bar{y}$-axis (note the bar!) followed by a reflection K'' in the $\bar{x}\bar{z}$-plane. That the result is the operation K is geometrically evident. Hence

$$Ku_m^l = R^{-1}(\mathrm{A}, \mathrm{B}, \Gamma) R(0, \pi, 0) K'' R(\mathrm{A}, \mathrm{B}, \Gamma) u_m^l$$

$$= {}^lD_{mn}^{-1}(\mathrm{A}, \mathrm{B}, \Gamma) {}^lD_{kn}(0, \pi, 0) f_n(\bar{z}_1, \bar{x}_2, \bar{z}_2, \bar{x}_3, -\bar{y}_3, \bar{z}_3 \cdots \bar{x}_N, -\bar{y}_N, \bar{z}_N).$$

As

$${}^lD_{kn}(0, \pi, 0) = (-1)^{l+k} \delta_{k,-n},$$

[34] The expression ${}^lD_{mn}^{-1}(0,\theta,\phi)\, f_n(r)$ is only apparently more general. This becomes obvious if it be recalled that the surface spherical harmonics form a complete orthogonal set in terms of which any single valued function of θ and ϕ may be expanded.

this is equivalent to

$$Ku^{l}_{m} = {}^{l}D^{-1}_{mn}(\mathrm{A}, \mathrm{B}, \Gamma)(-1)^{l+n} f_{-n}(\bar{z}_1, \bar{x}_2, \bar{z}_2, \bar{x}_3, -\bar{y}_3, \bar{z}_3 \cdots \bar{x}_N, -\bar{y}_N, \bar{z}_N).$$

Comparing this with Eq. (126), it is seen that if the arbitrary functions f_n are restricted by the relations

$$(-1)^{l+n} f_n(\bar{z}_1, \bar{x}_2, \bar{z}_2, \bar{x}_3, \bar{y}_3, \bar{z}_3 \cdots) = \pm f_{-n}(\bar{z}_1, \bar{x}_2, \bar{z}_2, \bar{x}_3, -\bar{y}_3, \bar{z}_3 \cdots),$$

then the second of Eq. (123) may be satisfied for either sign.

In case $N=1$, however, the situation is otherwise, since here $n=0$, and the equation is then

$$(-1)^{l} f(r) = \pm f(r),$$

which can only be satisfied if $\pm 1 = (-1)^{l}$. The sign is thus determined by l, is $+$ if l is even, $-$ if l is odd. This exceptional property for $N=1$ will be found to be responsible for one of the selection principles of the hydrogen-like atoms.

21. **The infinitesimal rotations and the angular momentum operators.** The analogy between Eqs. (124) and (126) will lead the reader to suspect that l and m have the same physical significance in both expressions, namely that of azimuthal and magnetic quantum members. That this is true will be shown by calculating the matrices which represent the angular momentum operators.[35] This is very conveniently accomplished by considering the infinitesimal rotations, which rotate a given point x, y, z, into another, x', y', z', which is only infinitesimally distant. It is obvious that the vector $x'-x, y'-y, z'-z$, must be perpendicular to x, y, z. There are three such rotations, one about each of the coordinate axes:

$$\begin{aligned}
R_x: \quad & x = x' \\
& y = y' + \epsilon z' \\
& z = z' - \epsilon y', \\
\\
R_y: \quad & x = x' - \epsilon z' \\
& y = y' \\
& z = z' + \epsilon x', \\
\\
R_z: \quad & x = x' + \epsilon y' \\
& y = y' - \epsilon x' \\
& z = z',
\end{aligned} \tag{128}$$

in which ϵ is a real infinitesimal. When one of these substitutions is performed on a function of the coordinates of the N electrons, the result is, e.g.,

[35] Weyl. *l.c.* Chapter IV a.

$$R_x\psi(x_i, y_i, z_i) = \psi(x_i, y_i + \epsilon z_i, z_i - \epsilon y_i)$$
$$= \psi(x_i, y_i, z_i) + \epsilon \sum_{i=1}^{N} \left(z_i \frac{\partial \psi}{\partial y_i} - y_i \frac{\partial \psi}{\partial z_i} \right). \tag{129}$$

The differential operators appearing in the sum will be familiar to the reader as (essentially) the operators associated to the angular momenta of the various electrons. If the operators associated[36] to the total angular momentum of the atom are designated by L_x, L_y, L_z, then

$$L_x\psi = -\frac{1}{i} \sum \left(y_i \frac{\partial \psi}{\partial z_i} - z_i \frac{\partial \psi}{\partial y_i} \right),$$
$$L_y\psi = -\frac{1}{i} \sum \left(z_i \frac{\partial \psi}{\partial x_i} - x_i \frac{\partial \psi}{\partial z_i} \right), \tag{130}$$
$$L_z\psi = -\frac{1}{i} \sum \left(x_i \frac{\partial \psi}{\partial y_i} - y_i \frac{\partial \psi}{\partial x_i} \right),$$

and the rotations R_x, R_y, R_z become

$$R_x = I + i\epsilon L_x,$$
$$R_y = I + i\epsilon L_y, \tag{131}$$
$$R_z = I + i\epsilon L_z.$$

It is a simple matter to calculate the matrices associated to R_x, R_y, R_z: the Cayley-Klein parameters are

$$R_x : a = 1, \qquad b = i\epsilon/2$$
$$R_y : a = 1, \qquad b = \epsilon/2$$
$$R_z : a = 1 + i\epsilon/2, \quad b = 0.$$

Giving $u_m{}^l$ the significance of Eq. (102) for the moment, and using Eq. (105), it is seen that

$$R_x u_m{}^l = u_m{}^l + \frac{i\epsilon}{2} \{ [(l + m)(l - m + 1)]^{1/2} u_{m-1}{}^l + [(l - m)(l + m + 1)]^{1/2} u_{m+1}{}^l \}$$

$$R_y u_m{}^l = u_m{}^l + \frac{\epsilon}{2} \{ [(l + m)(l - m + 1)]^{1/2} u_{m-1}{}^l - [(l - m)(l + m + 1)]^{1/2} u_{m+1}{}^l \}$$

$$R_z u_m{}^l = u_m{}^l (1 + im\epsilon),$$

whence

$$(L_x - iL_y) u_m{}^l = [(l - m)(l + m + 1)]^{1/2} u_{m+1}{}^l$$
$$(L_x + iL_y) u_m{}^l = [(l + m)(l - m + 1)]^{1/2} u_{m-1}{}^l \tag{132}$$
$$L_z u_m{}^l = m u_m{}^l,$$

[36] E. C. Kemble, Phys. Rev. Suppl. **1**, 194 (1929).

and

$$L^2 u_m{}^l = (L_x{}^2 + L_y{}^2 + L_z{}^2) u_m{}^l = l(l+1) u_m{}^l. \tag{132a}$$

These equations show that l and m are really the total angular momentum and its component in the z-direction, each measured in units of $h/2\pi$.

For half-integral values of l, which acquire a significance when the spin is taken into account, the angular momentum matrices may still be defined by equations similar to (131). In the case $l = 1/2$, they are

$$L_x = \tfrac{1}{2} \begin{Vmatrix} 0 & 1 \\ 1 & 0 \end{Vmatrix} \quad L_x = \tfrac{1}{2} \begin{Vmatrix} 0 & -i \\ i & 0 \end{Vmatrix}, \quad L_z = \tfrac{1}{2} \begin{Vmatrix} 1 & 0 \\ 0 & -1 \end{Vmatrix} \tag{133}$$

which are essentially the matrices which form the starting point of the Pauli spin theory.

22. **The vectorial addition of angular momenta as the reduction of a given manifold.** As a first example of the reduction of a given manifold, the case of a dynamical system which can be considered as composed of two parts loosely coupled together may be considered. This problem was treated with a high degree of success by Landé, who assigned an angular momentum to each part of the system. These were quantized and added together vectorially. The additional quantum condition that the resultant must also be integral or half integral was then introduced. This theory resulted in a rational classification of complex spectra. It will be shown that the Landé rules are obtained without difficulty by an application of the preceding methods.

The wave-equation of the system is in this case

$$(H_1 + H_2 + \epsilon H_{12}) \psi = W \psi, \tag{134}$$

where the operators H_1 and H_2 are functions of different sets of coordinates. The solutions of

$$H_1 u = W_1 u$$

and

$$H_2 v = W_2 v \tag{135}$$

are assumed to be known, and each of the Eqs. (134) and (135) are supposed to admit the operations of the rotation group as integrals. For a given value of W_1 there will then be $2l+1$ solutions $u_m{}^l$, and for W_2, $2\lambda+1$ solutions $v_\mu{}^\lambda$, each set of which is the basis of a manifold invariant under the rotations. There will be six angular momentum operators L_x, L_y, L_z and Λ_x, Λ_y, Λ_z; the total angular momenta are $J_x = L_x + \Lambda_x$, etc. It is to be noted that the L's and Λ's commute, and that L^2 and Λ^2 commute with J_x, J_y, J_z. These last remarks are not essential to the following considerations, but it will be convenient to use them in interpreting the results.

The zero approximations to the solutions of Eq. (134) will be linear combinations of the functions $u_m{}^l v_\mu{}^\lambda$. These products subtend a manifold of $(2l+1)$ $(2\lambda+1)$ dimensions which is obviously invariant under the rotation group, for

$$Ru_m{}^l v_\mu{}^\lambda = {}^lD_{mn}{}^\lambda D_{\mu\nu} u_m{}^l v_\nu{}^\lambda . \tag{136}$$

The matrices ${}^lD_{mn}{}^\lambda D_{\mu\nu}$ (whose rows are designated by the two indices m and μ, and whose columns by n and ν) are thus a (reducible) representation of the rotation group. The problem of reducing this representation will be completely solved when $(2l+1)$ $(2\lambda+1)$ functions

$$w^J_M = \sum_m \sum_\mu A^{J\,l\lambda}_{M\,m\mu} u_m{}^l v_\mu{}^\lambda \tag{137}$$

are determined,[37] such that

$$Rw^J_M = {}^JD_{MN} w^J_N . \tag{138}$$

The first part of this problem[38] is to determine the values of J for which such functions exist, and the number of independent ones for each J and M. The theory of characters (Section 10) enables these to be determined at once: the character of the representation (136) is

$$\chi(\omega) = \sum_{m=-l}^{l} \sum_{\mu=-\lambda}^{\lambda} e^{i(m+\mu)\omega}$$

(cf. Eq. (111)), which may be transformed into

$$\begin{aligned} \chi(\omega) &= \sum_{J=|l-\lambda|}^{l+\lambda} \sum_{M=-J}^{J} e^{iM\omega} \\ &= \sum_{J=|l-\lambda|}^{l+\lambda} {}^J\chi(\omega) \end{aligned} \tag{139}$$

where ${}^J\chi(\omega)$ are the primitive characters of the rotation group. Hence there is one and only one w^J_M for each of the values $J=l+\lambda$, $J=l+\lambda-1$, $\cdots$ $J=|l-\lambda|$ and none for any other values of J. This coincides with the result obtained by Landé.

It remains to determine the coefficients $A_{Mm\mu}{}^{J\,l\lambda}$ for these various values of J amd M. It is obvious that if this is accomplished for a particular set of functions $u_m{}^l\, v_\mu{}^\lambda$ satisfying Eq. (136), the same coefficients will apply to any other set satisfying this equation. For this reason it is permissable to take $u_m{}^l = \xi_1{}^{l+m}\eta_1{}^{l-m}/c(l,m)$ and $v_\mu{}^\lambda = \xi_2{}^{\lambda+\mu}\eta_2{}^{\lambda-\mu}/c(\lambda,\ \mu)$ where $\xi_1\eta_1$ and $\xi_2\eta_2$ are two pairs of variables which are both subjected to the substitution (79) and $c(l,\ m) = \sqrt{(l+m)!\ (l-m)!}$ The functions w^J_M will then be homogeneous polynomials in these four variables, of degree $2l$ in $\xi_1\eta_1$ and 2λ in $\xi_2\eta_2$. If such polynomials can be found which satisfy Eq. (138) for a given value of J, their coefficients will be, essentially, the $A_{Mm\mu}{}^{J\,l\lambda}$.

Now, if x is any arbitrary constant, the polynomials

$$f_M{}^J = (\xi_1 + x\xi_2)^{J+M}(\eta_1 + x\eta_2)^{J-M}/c(J,\ M) \tag{140}$$

[37] The coefficients A $\cdots$ have so many indices that it is expedient to exempt them all from the summation convention and to indicate all sums explicitly.

[38] The remaining considerations of this paragraph are purely mathematical and do not depend on J, l, λ being integers. They will thus still be applicable when the spin of the electron is taken into account (cf. Section 28) and R is replaced by D or RD.

obviously satisfy Eq. (138) for $\xi=\xi_1+x\xi_2$ and $\eta=\eta_1+x\eta_2$ undergo the same substitution as ξ_1, η_1 and ξ_2, η_2. They may be expanded in powers of x, and since x is arbitrary, the coefficients of x^p must also satisfy Eq. (138). Call these $f_M{}^{J}{}_{p}$; they are polynomials of degree $2J-p$ in $\xi_1\eta_1$ and degree p in $\xi_2\eta_2$. Furthermore, it is readily proven that

$$D(\alpha,\beta,\gamma)(\xi_1\eta_2-\xi_2\eta_1)^q=(\xi_1\eta_2-\xi_2\eta_1)^q; \tag{141}$$

hence the functions

$$w_M{}^{J}=Cf_M{}^{J}{}_{p}(\xi_1\eta_2-\xi_2\eta_1)^q, \tag{142}$$

where C is a constant independent of M, must satisfy Eq. (138); if q is a positive integer or zero, $w^{J}{}_{M}$ is a polynomial in the four variables and of degree $2J-p+q$ in $\xi_1\eta_1$ and $p+q$ in $\xi_2\eta_2$. If

$$p=J-(l-\lambda)$$

$$q=(l+\lambda)-J$$

it is therefore the polynomial required. Its coefficients can readily be calculated from Eqs. (140) and (142), and the A's from the equation

$$w_M{}^{J}=\sum_m\sum_\mu A^{Jl\lambda}_{Mm\mu}\frac{\xi_1{}^{l+m}\eta_1{}^{l-m}}{c(l,m)}\frac{\xi_2{}^{\lambda+\mu}\eta_2{}^{\lambda-\mu}}{c(\lambda,\mu)}. \tag{143}$$

This procedure yields precisely one set of A's for each value of J indicated by Eq. (139). The reduction of the basis $u^l_m v_\mu{}^\lambda$ is thus completed.[39]

As an example, let $\lambda=1$, $l>0$ then $J=l+1$, l, or $l-1$. For $J=l-1$, $p=0$, $q=2$ and

$$f_m{}^{l-1,0}=\xi_1{}^{l+m-1}\eta_1{}^{l-m-1}/c(l-1,m),$$

and therefore

$$w_m{}^{l-1}=C\xi_1{}^{l+m-1}\eta_1{}^{l-m-1}(\xi_1\eta_2-\xi_2\eta_1)^2/c(l-1,m),$$

or

$$w_m{}^{l-1}=\left[\frac{(l+m+1)(l+m)}{2l(2l+1)}\right]^{1/2}u^l{}_{m+1}v^1{}_{-1}-\left[\frac{(l+m)(l-m)}{l(2l+1)}\right]^{1/2}u^l{}_m v_0{}^1+\left[\frac{(l-m)(l-m+1)}{2l(2l+1)}\right]^{1/2}u^l{}_{m-1}v_1{}^1. \tag{144a}$$

The arbitrary constant C has been chosen so that the sum of the squares of the coefficients is unity (normalization). For $J=l$, $p=1$ and $q=1$, while for $J=l+1$, $p=2$ and $q=0$. The reader will verify that

[39] This procedure is used by Weyl, *l.c.* p. 159.

$$w_m{}^l = \left[\frac{(l-m)(l+m+1)}{2l(l+1)}\right]^{1/2} u^l{}_{m+1} v_{-1}{}^1 + \frac{m}{[l(l+1)]^{1/2}} u^l{}_m v_0{}^1 - \left[\frac{(l+m)(l-m+1)}{2l(l+1)}\right]^{1/2} u_m{}^l{}_{-1} v_1{}^1$$

$$w_m{}^{l+1} = \left[\frac{(l-m+1)(l-m)}{2(l+1)(2l+1)}\right]^{1/2} u^l{}_{m+1} v_{-1}{}^1 + \left[\frac{(l+m+1)(l-m+1)}{(l+1)(2l+1)}\right]^{1/2} u^l{}_m v_0{}^1 + \left[\frac{(l+m+1)(l+m)}{2(l+1)(2l+1)}\right]^{1/2} u^l{}_{m-1} v_1{}^1. \tag{144b}$$

In the exceptional case, $l=0$, which was excluded above, the result is obviously $w_m{}^1 = u^0 v_m{}^1$.

The A's have numerous properties, which may be deduced from their definition. Since ${}^J D_{MN}$ is a unitary representation of the group, it follows from Section 11, Theor. I that the constant C may always be chosen so that

$$(w_M{}^J, w_N{}^K) = \delta_{MN}\delta_{JK};$$

it may also be assumed that

$$(u^l{}_m v_\mu{}^\lambda, u^l{}_n v_\nu{}^\lambda) = \delta_{mn}\delta_{\mu\nu}.$$

It then follows from Eq. (137) that

$$\sum_m \sum_\mu A^{Jl\lambda}_{Mm\mu} A^{Kl\lambda}_{Nm\mu} = \delta_{NM}\delta_{JK}, \qquad \sum_J \sum_M A^{Jl\lambda}_{Mm\mu} A^{Jl\lambda}_{Mn\nu} = \delta_{mn}\delta_{\mu\nu}; \tag{145}$$

the transformation (137) is thus orthogonal.

A large number of the coefficients are zero. Since ${}^l D_{mn}(\omega, 0, 0) = e^{im\omega}\delta_{nm}$, etc., it follows from Eq. (137) and (138) that

$$A^{Jl\lambda}_{Mm\mu} = 0 \text{ unless } m + \mu = M. \tag{146}$$

From this, it follows: that $w_{l+\lambda}{}^{l+\lambda}$ contains only one term, viz., $u^l_l v_\lambda{}^\lambda$; that $w^J_{l+\lambda-1}$ contains only two (for $J = l+\lambda$, and $l+\lambda-1$; it vanishes for all other J); etc. etc. This result can most conveniently be summarized by tabulating the A's in matrix array (Fig. 4). Each row of this matrix is characterised by the two indices M and J; they may conveniently be arranged so that rows having the same M are adjacent and $M-1$ is lower than M in the series. Inside these sets of rows, J varies from $l+\lambda$ to M, say again in descending order. Each column of the matrix is also determined by two indices, m and μ (for l and λ are fixed). These columns may be arranged so that all having the same $m+\mu = M$ are adjacent, and $M-1$ is farther to the right than M. Within these sets, $m-1$ may always be placed farther to the right than m. When the A's are arrayed in this manner, all entries not in the outlined squares of the figure are zero; the figure is drawn for $l=2$, $\lambda=1$. The first

square on the diagonal contains only one entry; the second, two rows and two columns; the third three rows and three columns. This increase by one continues until the square contains either $2l+1$ or $2\lambda+1$ rows and columns, then the size remains constant for a time, after which a decrease by one sets in, in such a way that the last square contains just one entry again. Since the matrix as a whole is orthogonal (Eq. (138)), it follows that each of the smaller matrices into which it reduces is also orthogonal. This result is of importance for the calculation of intensities and energy values by the method of sums.[40]

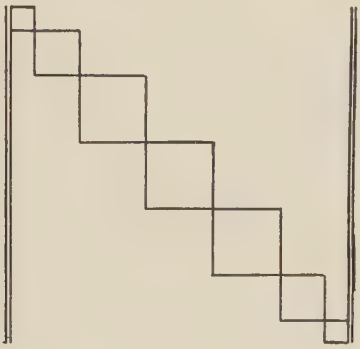


Fig. 4

From Eq. (137) and the properties of the operators L^2, Λ^2, J^2, etc., it follows that

$$\begin{aligned} L^2 w_M{}^J &= l(l+1) w_M{}^J, \\ \Lambda^2 w_M{}^J &= \lambda(\lambda+1) w_M{}^J, \\ J^2 w_M{}^J &= J(J+1) w_M{}^J, \\ J_z w_M{}^J &= M w_M{}^J. \end{aligned} \tag{147}$$

The matrices associated to these four operators by the functions $w_M{}^J$ are thus diagonal, a result which is often of importance. Interpreted in terms of the Landé theory, it means that the magnitudes of the vectors L and Λ remain constant in time (to this approximation) as does the magnitude of their resultant J. The separate components of the resultant are also a constant, but this is not true of the components of L and Λ. These vectors are to be considered as precessing about J as an axis. Since

$$L_x{}^2 = (J_x - \Lambda_x)^2, \text{ etc.},$$

it follows that

$$L^2 = J^2 - 2(J_x\Lambda_x + J_y\Lambda_y + J_z\Lambda_z) + \Lambda^2$$

whence it follows from (147) that

$$2(J_x\Lambda_x + J_y\Lambda_y + J_z\Lambda_z) w_M{}^J = [J(J+1) + \lambda(\lambda+1) - l(l+1)] w_M{}^J \tag{148}$$

[40] Cf. Sections 23 and 31.

which shows that the matrix associated to this operator is also diagonal. This means that the projection of Λ on J is a constant, in accordance with the last remark above. Similarly

$$2(L_x\Lambda_x + L_y\Lambda_y + L_z\Lambda_z)w_M{}^J = [l(l+1) + \lambda(\lambda+1) - J(J+1)]w_M{}^J. \quad (148a)$$

If the functions $u_m{}^l$ subtend the whole of the manifold of W_1 and $v_\mu{}^\lambda$ the whole manifold of W_2, so that these energy levels are just $(2l+1)$- and $(2\lambda+1)$-fold degenerate, the perturbation matrix

$$(H_{12}w_M{}^J, w_N{}^K)$$

is diagonal also. This follows at once from the considerations of Section 6, since there is no accidental degeneracy in J. If further degeneracy is present, this is not necessarily the case, but the non-diagonal elements will be relatively few in number and arranged as explained in that section.

23. **The method of sums.** The foregoing methods are capable of extension to the case of a given manifold subtended by a basis whose functions are products of, say N factors each of the form $u_m{}^l$. The manifold is certainly invariant under the rotation group and its basis can therefore be reduced to the form $w_M{}^{L\alpha}$ $(\alpha = 1 \cdots q(L))$. The values of L and the range of the index α which distinguishes different irreducible sub-bases of the same character may both be determined from the character of the representation. A slight generalization of the method of Eqs. (140)–(143) furnishes the coefficients of the transformation.

These calculations will not be carried out in detail, however, but another method which often enables the calculation of the characteristic values of the perturbation matrix without solving the secular equation will be explained.[41] Let the functions of the given basis be $v_m{}^k$ where m is the sum of the magnetic quantum numbers of the N factors and the index k is assigned in any manner which serves to distinguish functions having the same m. In other words, the explicit form of the functions is not to be utilized in the following considerations, but only the two properties:

1. The manifold subtended by $v_m{}^k$ is invariant under the group $R(\alpha\beta\gamma)$;

2. $$R(\omega, 0, 0)v_m{}^k = e^{im\omega}v_m{}^k.$$

Then, because of (1), it is known that a transformation matrix can be found such that

$$w_M{}^{L\alpha} = a_{ML\alpha;km}v_m{}^k, \qquad (\alpha = 1 \cdots q(L)),$$

where

$$Rw_M{}^{L\alpha} = {}^LD_{MN}w_N{}^{L\alpha}.$$

From this and (2), it follows that $w_M{}^{L\alpha}$ can be a function only of those $v_m{}^k$ for which $m = M$:

$$w_M{}^{L\alpha} = a_{L\alpha;k}(M)v_M{}^k. \quad (149)$$

[41] J. C. Slater, Phys. Rev. **34**, 1293 (1929).

In this equation, L must necessarily be greater than or at most equal to M, but not greater than the greatest value of m; whatever other restrictions it may be subject to need not be considered.

Now, let $H_{km;jn}$ be the perturbation matrix in terms of the $v_m{}^k$, and let it be known to commute with the representation of the rotation group. Then it follows from the property (2) that

$$H_{km;jn} = H_{kj}(m)\delta_{mn}. \tag{150}$$

On transforming this with the matrix $a_{L\alpha;k}(M)$, it must assume the form[42]

$$H'_{\alpha\beta}(L)\delta_{LK}\delta_{MN}$$

where $H'_{\alpha\beta}(L)$ is independent of M. It will be no loss of generality if we assume

$$H'_{\alpha\beta}(L) = H(L, \alpha)\delta_{\alpha\beta}, \tag{151}$$

for if this is not the case, a second canonic transformation, obtained by solving the perturbation equations, will always bring it to this form. The quantities $H(L, \alpha)$ are the characteristic values which are to be calculated. Now, for any given M, the matrix $H(L,\alpha)\ \delta\alpha\beta\delta_{Lk}$ $(L \geqq M)$ must be equivalent to the matrix $H_{kj}(M)$ because of (149). Because of the invariance of the diagonal sum, therefore,

$$\sum_{L\geqq M}\ \sum_{\alpha=1}^{q(L)} H(L, \alpha) = \sum_{k} H_{kk}(M). \tag{152}$$

Since $H(L, \alpha)$ is independent of M, the result of subtracting the analogous equation for $M+1$ from Eq. (152) (and then replacing M by L) is

$$\sum_{\alpha=1}^{q(L)} H(L, \alpha) = \sum_{k} H_{kk}(L) - \sum_{k} H_{kk}(L+1). \tag{153}$$

This equation permits the calculation of the sum of all energy levels associated to a given value of L. In two particular cases, viz.: when $q(L)=1$ or when the dependence of $H(L\alpha)$ on α is known from other considerations, it permits the calculation of the energy levels themselves *without the solution of the secular equation.* The importance of the method is derived from the fact that a large number of actual cases are included under the two special cases just mentioned.

24. **Selection principles.** The quantities $A_{Mm\mu}{}^{Jl\lambda}$ also govern the selection principles and intensities of spectral lines. For the first, consider an atom whose characteristic functions for the energy W are $u_m{}^l$, for the energy W' are[43] $u'_\mu{}^\lambda$. Then a virtual oscillator is associated to the transition $W, l, m \rightarrow W', \lambda, \mu$ whose vector amplitude is given by the three quantities

[42] It is suggested that the reader compare these considerations with the diagrams of Section 6.

[43] These are now characteristic functions of the *same* system, not as in Section 22, of *different* partial systems.

$$\begin{matrix}(Xu_m{}^l, u'_\mu{}^\lambda)\\ (Yu_m{}^l, u'_\mu{}^\lambda)\\ (Zu_m{}^l, u'_\mu{}^\lambda)\end{matrix} \tag{154}$$

where

$$X = \sum_i x_i, \quad Y = \sum_i y_i, \quad Z = \sum_i z_i \tag{155}$$

The expressions (154) can be partially evaluated in terms of the $A_{Mm\mu}{}^{J\,l}$. To see this, note that the three functions

$$\begin{aligned} v_1{}^1 &= \frac{1}{2^{1/2}}(-X + iY)\\ v_0{}^1 &= Z\\ v_{-1}{}^1 &= \frac{1}{2^{1/2}}(X + iY), \end{aligned} \tag{156}$$

satisfy the relation

$$Rv_p{}^1 = {}^1D_{pq}v_q{}^1. \tag{157}$$

The expressions (154) will be known if

$$(v^1{}_p u_m{}^l, u'_\mu{}^\lambda)$$

is known for each p. Now, by means of Eq. (144), functions $w_M{}^J$ can be calculated in terms of the products $v^1{}_p u^l{}_m$, and conversely, the products can be expressed in terms of $w_M{}^J$:

$$v_p{}^1 u_m{}^l = \sum_{J=l-1}^{l+1} A^{J\,l1}_{m+p\ m\ p} w^J_{m+p}.$$

Hence

$$(v^1{}_p u^l{}_m, u_\mu{}'^\lambda) = \sum_{J=l-1}^{l+1} A^{J\,l1}_{m+p\ m\ p}(w^J_{m+p}, u'_\mu{}^\lambda)$$

As was shown in Section 11,

$$(w_M{}^J, u'_\mu{}^\lambda) = a(J)\delta_{J\lambda}\delta_{M\mu},$$

where a is independent of μ. Hence

$$(v'_p u_m{}^l, u'_\mu{}^\lambda) = a(l\lambda)A^{\lambda l1}_{\mu m p}$$

if $\mu = m+p$, $|l-\lambda| \leqq 1$, and zero otherwise.

This result does not depend at all on the explicit form of the functions X, Y, Z, but merely on the fact that they are components of a vector and, as such, experience the same substitution as the coordinates of a point. They might even have been vector operators, i.e., operators which convert a scalar function into a vector. The matrices representing any vector quantity

all have the following elements, which are readily determined by picking out the proper A's from Eq. (144):

$$\lambda = l + 1$$

$$\mu = m + 1,\ \frac{1}{2^{1/2}}(-X + iY):\ a(l, l+1)\left[\frac{(l+m+2)(l+m+1)}{2(l+1)(2l+1)}\right]^{1/2};$$

$$\mu = m,\qquad Z:\ a(l, l+1)\left[\frac{(l+m+1)(l-m+1)}{(l+1)(2l+1)}\right]^{1/2}; \tag{159}$$

$$\mu = m - 1,\ \frac{1}{2^{1/2}}(X + iY)\quad :a(l, l+1)\left[\frac{(l-m+2)(l-m+1)}{2(l+1)(2l+1)}\right]^{1/2};$$

$$\lambda = l$$

$$\mu = m + 1,\ \frac{1}{2^{1/2}}(-X + iY):\quad -a(l, l)\left[\frac{(l+m+1)(l-m)}{2l(l+1)}\right]^{1/2};$$

$$\mu = m,\qquad Z:\qquad a(l, l)\frac{m}{[l(l+1)]^{1/2}}; \tag{160}$$

$$\mu = m - 1,\ \frac{1}{2^{1/2}}(X + iY)\quad :\qquad a(l, l)\left[\frac{(l-m+1)(l+m)}{2l(l+1)}\right]^{1/2};$$

$$\lambda = l - 1$$

$$\mu = m + 1,\ \frac{1}{2^{1/2}}(-X - iY):\ a(l, l-1)\left[\frac{(l-m-1)(l-n)}{2l(2l+1)}\right]^{1/2};$$

$$\mu = m,\qquad Z: -a(l, l-1)\left[\frac{(l+m)(l-m)}{l(2l+1)}\right]^{1/2}; \tag{161}$$

$$\mu = m - 1,\ \frac{1}{2^{1/2}}(X + iY)\quad :\ a(l, l-1)\left[\frac{(l+m)(l+m-1)}{2l(2l+1)}\right]^{1/2};$$

all elements not explicitly indicated are zero. The angular momentum operators are vectors, and their matrices have this form, with $a(l, l+1) = a(l, l-1) = 0$ and $a(l, l) = [l(l+1)]^{1/2}$ (cf. Eq. (132)).

The probability of the transition $W, l, m \to W', \lambda, \mu$ is proportional to the sum of the absolute squares of the matrix elements XYZ associated to it. The reader will readily verify that the sum of the probabilities for a transition from W, l, m to $W', \lambda, m+1$, W', λ, m, and $W', \lambda, m-1$ is given by

$$|\, a(l, \lambda)\,|^2 \frac{2\lambda + 1}{2l + 1} \tag{162a}$$

and is therefore *independent of* m. A similar formula holds for the sum of the transition probabilities from $W, l, \mu-1$, W, l, μ, and $W, l, \mu+1$ to W', λ, μ: it is equal to

$$|\, a(l, \lambda)\,|^2. \tag{162b}$$

The Eqs. (159), (160), and (161) embody the selection principles for the total angular momentum l and its component along the z-axis m. There is a further selection principle arising from the presence of the operation K in the group $\mathfrak{G}$: since X, Y, Z are odd functions, $(KX = -X)$, it follows that

$$(v^1{}_p u_m{}^l, u'_\mu{}^\lambda) = -(v^1{}_p K u_m{}^l, K u'_\mu{}^\lambda);$$

hence the expression will vanish if $u_m{}^l$ and $u'_\mu{}^\lambda$ are both odd or both even. The result is Laporte's rule: *even terms combine only with odd terms and odd only with even.* In the case of the hydrogen atom, all terms with even l are even, and all terms with odd l are odd (cf. Section 20). The result is that $a(l, l) = 0$ in Eq. (160) so that $|\lambda - l| = 1$ for the hydrogen atom. How this rule applies in the case of more than one electron, will be seen in a later section.

It is to be noted that all elements (160) vanish for $l = \lambda = 0$, corresponding to the well-known special case of the selection principles.

25. **The addition of spin-vectors as the reduction of a manifold. The alternation rule for even and odd multiplets.** It is now necessary to consider the influence of the spin on the wave functions of an atom. The wave equation may be assumed to have the form

$$\{H(x_i y_i z_i) + H'(x_i y_i z_i s_i)\}\psi = W\psi. \tag{163}$$

It is indicated by experiment that the operator H' may offen be treated as a perturbation. If it be entirely neglected, the wave-function is

$$\psi(x_i, y_i, z_i, s_i) = u(x_i, y_i, z_i)\Xi(s_i) \tag{164}$$

where

$$Hu(x_i, y_i, z_i) = Wu(x_i, y_i, z_i) \tag{165}$$

and $\Xi(s_i)$ is any function of the spin variables.

The Eq. (163) is invariant under the group RD, and Eq. (165) under the group R. Hence the manifold of solutions u may be assumed to be completely reduced under this group: $u = u_m{}^l(x_i, y_i, z_i)$. Furthermore, the manifold of functions $\Xi(s_i)$ is invariant under the group of operations D. This manifold is 2^N dimensional (cf. Eq. (89)) and the functions

$$\begin{aligned} &\xi(s_1)\xi(s_2)\cdots\xi(s_N) \\ &\eta(s_1)\xi(s_2)\cdots\xi(s_N) \\ &\cdots\cdots\cdots\cdots \end{aligned} \tag{166}$$

may be taken as its basis. The problem of reducing this manifold is the one to be solved before any others are considered.[44] For typographical reasons, we introduce the notation

$$\begin{aligned} \xi_i &= \xi(s_i), \\ \eta_i &= \eta(s_i); \end{aligned} \tag{167}$$

[44] J. v. Neumann and E. Wigner, Zeits. f. Physik **47**, 200 (1928).

the problem is to find a polynomial $\Xi_\mu{}^S$ in the $2N$ quantities ξ_i, η_i, which shall be of first degree in each pair having a given index, and satisfy the equation

$$D\Xi_\mu{}^S = {}^SD_{\mu\nu}\Xi_\nu{}^S. \tag{168}$$

The values of S for which this problem possesses a solution and the number of independent solutions may be determined from the theory of characters. Since

$$D(\omega, 0, 0)\xi_i = e^{i\omega/2}\xi_i,\ D(\omega, 0, 0)\eta_i = e^{-i\omega/2}\eta_i,$$

the character of the representation based on the functions (166) is

$$\chi(\omega) = (2\cos\omega/2)^N.$$

Hence $q(S\ N)$, the number of times the irreducible representation ${}^SD_{\mu\nu}$ appears in the completely reduced form of this representation is (cf. Section 10 and Eq. (113)).

$$\begin{aligned} q(S, N) &= \frac{1}{4\pi}\int_0^{4\pi}(2\cos\omega/2)^N \sin\left[(2S+1)\omega/2\right]\sin\left[\omega/2\right]d\omega \\ &= \binom{N}{N/2 - S} - \binom{N}{N/2 - S - 1}, \end{aligned} \tag{169}$$

in which ( ) is the binomial coefficient. The values of S for which the problem possesses a solution are therefore

$$S = \frac{N}{2},\ \frac{N}{2} - 1,\ \frac{N}{2} - 2 \cdots \left(\frac{1}{2}\text{ or }0\right). \tag{170}$$

The values of S are integral or half integral according as N is an even or odd integer. This will prove to be the alternation rule for the occurrence of multiplets in the periodic system of elements.

When S has one of the values (170), the problem possesses $q(S, N)$ independent solutions. This function is seen to satisfy the recursion formula

$$q(S, N) = q(S + \tfrac{1}{2}, N - 1) + q(S - \tfrac{1}{2}, N - 1), \tag{171}$$

which corresponds to the earlier ideas regarding the addition of spin vectors: for $N=2$, the two vectors may either be parallel or antiparallel, resulting in $S=1$ or $S=0$; for $N=3$, $S=3/2$ may be realized in one way: by adding the third s parallel to two which are already parallel, while the value $S=1/2$ may be realized in either of two ways: by adding the third antiparallel to two which are parallel, or by adding it to two in the antiparallel orientation. If this process of successive addition be followed through, it will be found that the Eq. (171) is always satisfied. The values of the function $q(S, N)$ are given in Table I.

Polynomials satisfying Eq. (168) are readily constructed: if $a_1, a_2, \cdots a_{2S-1}$, a_{2S} are any $2S$ constants, the function

$$(a_1\xi_1 + \cdots + a_{2S}\xi_{2S})^{S+\mu}(a_1\eta_1 + \cdots + a_{2S}\eta_{2S})^{S-\mu}/c(S, \mu) \tag{172}$$

TABLE I

S \ N	0	1	2	3	4	5	6	7	8	9
0	1		1		2		5		14	
1/2		1		2		5		14		42
1			1		3		9		28	
3/2				1		4		14		48
2					1		5		20	
5/2						1		6		27
3							1		7	
7/2								1		8
4									1	
9/2										1

is obviously one of such a set, as is the coefficient of $a_1\, a_2 \cdots a_{2S}$. Call this $f_{\mu S}$; it is linear in each of the $2S$ pairs $\xi_1, \eta_1; \cdots \xi_{2S}, \eta_{2S}$. A polynomial which is of first degree in all pairs is

$$\Xi_{\mu S} = f_\mu{}^S(\xi_{2S+1}\eta_{2S+2} - \xi_{2S+2}\eta_{2S+1}) \cdots (\xi_{N-1}\eta_N - \xi_N\eta_{N-1}). \qquad (17\mathfrak{E})$$

Other polynomials of this kind may be obtained by permuting the indices of the quantities ξ_i, η_i. There are $N![/(2S)!\,(N/2-S)!\,2^{N/2-S}]$ *different* polynomials which arise in this way, but only $q(S, N)$ of them can be linearly independent. These will be indicated by a third index, p, which runs from $1 \to q(S, N)$: $\Xi_\mu{}^{Sp}$.

As an example, let it be required to find the irreducible submanifolds $S=1$, for $N=4$. They are three in number, since $q(1, 4)=3$. If $\mu=1$, the expression (172) becomes

$$(a_1\xi_1 + a_2\xi_2)^2/2^{1/2}.$$

The coefficient of a_1a_2 in this is $2^{1/2}\xi_1\xi_2$ and the function

$$\Xi_1{}^{11} = 2^{1/2}\xi_1\xi_2(\xi_3\eta_4 - \xi_4\eta_3);$$

the other functions may be taken to be

$$\Xi_1{}^{12} = 2^{1/2}\xi_2\xi_3(\xi_4\eta_1 - \xi_1\eta_4),$$

$$\Xi_1{}^{13} = 2^{1/2}\xi_3\xi_4(\xi_1\eta_2 - \xi_2\eta_1).$$

For $\mu=0$, the expression (172) is

$$(a_1\xi_1 + a_2\xi_2)(a_1\eta_1 + a_2\eta_2)$$

in which the coefficient of a_1a_2 is

$$(\xi_1\eta_2 + \xi_2\eta_1);$$

hence

$$\Xi_0{}^{11} = (\xi_1\eta_2 + \xi_2\eta_1)(\xi_3\eta_4 - \xi_4\eta_3),$$

$$\Xi_0{}^{12} = (\xi_2\eta_3 + \xi_3\eta_2)(\xi_4\eta_1 - \xi_1\eta_4),$$

$$\Xi_0{}^{13} = (\xi_3\eta_4 + \xi_4\eta_3)(\xi_1\eta_2 - \xi_2\eta_1).$$

The functions for $\mu = -1$ may be obtained from those for $\mu = 1$ by replacing ξ by η in the first two factors of each term.

The basis of the manifold of the zero approximations to the solutions of Eq. (163) has therefore been brought to the form

$$u^l{}_m \Xi_\mu{}^{Sp}$$

which contains $(2l+1)\ 2^N$ functions. It is invariant under the operations RD:

$$RDu_m{}^l \Xi_\mu{}^{Sp} = {}^lD_{mn}{}^S D_{\mu\nu} u_n{}^l \Xi_\nu{}^{Sp}. \tag{174}$$

The problem of completely reducing this manifold may readily be solved by the method of Section 22. If

$$w_M{}^{Jp} = \sum_m \sum_\mu A^{Jl^S}_{M m\mu} u^l{}_m \Xi_\mu{}^{Sp}$$

then

$$RDw_M{}^{Jp} = {}^JD_{MN} w_N{}^{Jp}. \tag{175}$$

which shows that the complete reduction has been accomplished. J is the inner quantum number, which represents the total angular momentum of the atom.

It remains to consider the action of the operator K; obviously

$$Kw_M{}^{Jp} = \pm\ w_M{}^{Jp}$$

according as

$$Ku_m{}^l = \pm\ u_m{}^l.$$

There is thus no general relation between the value of J and the odd or even character of $w_M{}^{Jp}$

The effect of the perturbation H' will be to remove the degeneracy in p. It will not be able to remove the degeneracy in M, so that to *every approximation*, the energy level will be characterized by a certain value of J and $2J+1$ values of M. The selection principles for J and M are readily derived, for the probability of the transition from $u_M{}^J$ to $u'_N{}^L$ is determined by $(v_\mu^1 u_M{}^J, u'_N{}^L)$ where $v_\mu{}^1$ is still given by Eq. (156) and is independent of s_i. Since $RDv_\mu{}^1 = {}^1D_{\mu\nu} v_\nu{}^1$, all the reasoning of Section 24 is still valid and the selection principles are therefore

$$\begin{gathered} \Delta J = \pm\ 1 \text{ or } 0 \\ \Delta M = \pm\ 1 \text{ or } 0 \\ \text{odd} \rightleftarrows \text{even} \\ J = 0 \text{ does not combine with } J = 0 \end{gathered} \tag{176}$$

and the intensities of permitted transitions are given by Eqs. (159) – (161), with appropriate changes of notation.

V. The Permutation Group and the Pauli Exclusion Principle

26. **Physical considerations regarding the necessity and possibility of an exclusion principle.** It must be accepted as one of the most fundamental conclusions of modern physics that all electrons are identical. They have no distinguishing features, as have the planets of the solar system; each has identically the same charge, the same mass,—in short, when two electrons are placed in identical environments, they are acted on by the same forces and behave in the same manner. This fact may be expressed in terms of the wave-function: according to the quantum theory, $\psi\psi^*$ is the distribution function of a statistical ensemble of dynamical systems. If each of these contains N electrons, then $\psi' = P\psi$ (cf. Section 13) represents another ensemble differing from the first only in that the electrons have exchanged places. This second ensemble must be physically indistinguishable from the first ensemble.

For the purposes of the following pages it will be very convenient to denote the four variables of the i-th electron by the single symbol i. The wave function is then $\psi(1, 2, 3, \cdots N)$ and the wave equation is

$$H(1, 2, 3, \cdots N)\psi = W\psi. \tag{177}$$

The first consequence of the remarks of the preceding paragraph is that

$$H(1, 2, 3 \cdots N) = H(2, 1, 3 \cdots N) = \text{etc., etc.} \tag{178}$$

The operator may be said to be symmetric in the variables $1, 2, 3, \cdots N$, and

$$HP = PH. \tag{179}$$

The operations P are members of the group $\mathfrak{G}$ of the Eq. (177) as we have already seen. This is all that can be deduced from Eq. (177), but the consideration that the ensemble represented by ψ' must be indistinguishable from the ensemble represented by ψ results in more. The general principles of quantum-mechanics show that it can only be true if

$$\psi' = c\psi$$

where

$$cc^* = 1. \tag{180}$$

Solutions of Eq. (177) not having this property are therefore physically inadmissable. This enables an *a priori* restriction of the representations of the permutation group which need be considered: only those which are "matrices of one row and column" are required.

It can be shown that there are only two of these, and of them, one (the so-called "symmetric") does not correspond to any of the energy levels which are actually observed in atomic spectra. Only the "antisymmetric" one will therefore come into our consideration: the other governs the theory of photons, and may also be involved molecular spectra. Despite this fact, it will occasionally be convenient to consider other representations of the permutation group.

Before proceding to a study of these, it is well to convince ourselves that this "exclusion principle" is really consistent with the general principles of quantum dynamics. Is it not possible that if an atom is in a state represented by a function satisfying (180), it may, either spontaneously or otherwise, jump to a state which does not satisfy this condition? This may be answered in the negative, for first of all there can be no forced transitions to such a state for the perturbation causing a forced transition will be represented in Eq. (177) by an additional *symmetric* operator, H′ $(1 \cdots N)$. As the forbidden state will belong to the basis of an irreducible representation which is inequivalent to the permitted state, there will be no terms in the perturbation matrix corresponding to such a transition. The same reasoning leads to the conclusion that there will be no spontaneous transitions, for the functions X, Y, Z, (Eq. (155)) are also symmetric functions. The mathematical scheme is thus entirely consistent with the imposition of the additional exclusion principle.

27. **The permutation group and its representations.** The operation of replacing the natural sequence of objects $123 \cdots N$ by the sequence $i_1 i_2 \cdots i_N$ can be considered as the result of a number of interchanges. By an interchange is meant the operation which replaces the element i by the element j, and j by i leaving all others unmoved. Thus the sequence 1342 can be obtained from 1234 by first interchanging 2 and 3, then 2 and 4. This factorization of a given permutation into interchanges can be accomplished in more than one way. The number of interchanges required for one factorization may even be different than the number required for another; however the number will always be $\left\{\begin{matrix}\text{odd}\\ \text{even}\end{matrix}\right\}$ if it is $\left\{\begin{matrix}\text{odd}\\ \text{even}\end{matrix}\right\}$ in one particular case. Thus all permutations may be divided into two classes, odd and even. This classification is the basis of the theory of determinants.[45] It is seen that the product of two odd or two even permutations is even, the product of an odd and an even permutation is an odd permutation. The two one-dimensional representations of the permutation group are now readily discovered: the symmetric one is

$$P\psi = \psi, \tag{181a}$$

the antisymmetric

$$P\psi = \epsilon(P)\psi \tag{181b}$$

where $\epsilon(P) = +1$ if P is an even permutation, and $\epsilon(P) = -1$ if P is odd.

The antisymmetric representation is the only one which will be fundamental in the following. It will be convenient to know, however, that the functions Ξ_μ^{Sp} (Section 25) constitute a basis for an[46] irreducible representation:

[45] For a convenient summary of the theory of determinants, cf. "Mathematical Monographs," No. 3. by L. G. Weld (*Ed.* Merriman and Woodward) Wiley & Sons. (1906).

[46] The representations ${}^{S}P_{pp}$, $S=0, 1/2 \cdots$, do not constitute a complete set of representations of the permutation group; there are others which cannot be brought into this form. Cf J. v. Neumann and E. Wigner, Zeits f. Physik **49**, 80 (1928), F. Hund, Zeits. f. Physik **43**, 788 (1927).

$$P\Xi_\mu{}^{Sp} = {}^{S}P_{pq}\Xi_\mu{}^{Sq}. \tag{182}$$

That the manifold of functions subtended by $\Xi_\mu{}^{Sp}$ is invariant under the permutations is obvious from its construction; that it is an irreducible manifold will not be proven. The matrices ${}^{S}P_{pq}$ are independent of μ, but dependent on S; the first statement follows at once from the fact that $PD=DP$. The elements ${}^{S}P_{pq}$ are real numbers.

Furthermore, if P_{nm} be any representation of the P's, so is

$$'P_{nm} = \epsilon(P)P_{nm}, \tag{183}$$

as follows at once from $\epsilon(P)\,\epsilon(Q)=\epsilon(PQ)$. If P_{nm} is irreducible, so is $'P_{nm}$, as may be seen at once from Schur's lemma. The two characters are related by

$$'\chi(P) = \epsilon(P)\chi(P); \tag{184}$$

the two irreducible representations can therefore be equivalent only if $\chi(P)=0$ for all odd permutations—a case which does occur. The representation $'P_{nm}$ is said to be reciprocal to P_{nm}. These are the only theorems regarding the representations of the permutation group which are required for the following.

28. **The Pauli exclusion principle and normal multiplets.** The Pauli exclusion principle was derived from a detailed consideration of the empirical material obtained from observation of the spectra of complex atoms. It states that in a given atom, no two electrons are in the same state. Two electrons may be in states having the same principal, azimuthal, and magnetic quantum numbers, but then their spin vectors must be oppositely directed. If their spins are parallel, at least one of the three other quantum numbers must be different. No real exception to this rule has ever been observed. It may be shown that this principle is equivalent to excluding the symmetric representation and retaining the antisymmetric representation as the only one having any physical significance.

To see this, it is convenient to consider the zero approximation to the wave equation, which neglects all interaction between electrons, considering only forces exerted on them by the nucleus. To this approximation the wave equation is

$$\sum_{i=1}^{N} H(i)\psi = W\psi, \tag{185}$$

where $H(i)$ is an operator depending only on the four coordinates of the ith electron. It is assumed that the solutions and characteristic values of

$$H(1)u(1) = Wu(1) \tag{186}$$

are known. Then a solution of Eq. (185) is

$$\psi(1, 2, 3, \cdots N) = u_1(1)u_2(2)u_3(3)\cdots u_N(N) \tag{187}$$

where $u_1, u_2, \cdots u_N$ are N solutions of the Eq. (186). The energy level W

is the sum $W_1+W_2+\cdots+W_N$ of the characteristic values to which u_1, $u_2, \cdots u_N$ belong. It will, for the moment, be assumed that these are not degenerate.

Then there will be at most $N!$ independent solutions of Eq. (185) for this value of W, which are obtainable from (187) by permuting the variables entering into the various u's: thus

$$\psi(3, 6, 2 \cdots n) = u_1(3)u_2(6)u_3(2) \cdots u_N(n)$$

is also a solution. From the manifold subtended by these functions, it is possible to select that function which satisfies Eq. (181a) or (181b) by the general method of Section 9. Thus

$$v = \sum_P P\psi(1, 2, \cdots N) \tag{188}$$

satisfies Eq. (181a) and

$$w = \sum_P \epsilon(P)P\psi(1, 2, \cdots N) \tag{189}$$

satisfies Eq. (181b). This last may also be written as a determinant:

$$w = \begin{vmatrix} u_1(1) & u_2(1) & \cdots & u_N(1) \\ u_1(2) & u_2(2) & \cdots & u_N(2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ u_1(N) & u_2(N) & \cdots & u_N(N) \end{vmatrix}. \tag{190}$$

The Pauli exclusion principle states that if $u_1(1) = u_2(1)$ the state of the atom represented by v or w can have no physical significance. This is automatically fulfilled in the case of w, for in this case the determinant (190) will have two identical rows, and hence $w=0$. The symmetric function v will still exist, however, even in this case. If the atom were ever in a state represented by a symmetric function but obeying the Pauli exclusion principle, it would be possible for it to perform a transition to a state such as has just been described which does not obey the exclusion principle. The conclusion must therefore be that the state of atom is never represented by a symmetric wave-function. This result applies immediately only to the zero-approximation but there is no question but that it is the precise and general formulation of the exclusion principle.

It is then necessary to consider the effect of the interaction of the electrons on the antisymmetric wave-function only. This problem may be attacked in several ways, the most obvious being to consider the determinant (190) as the zero approximation to the solution of the complete wave equation and to treat the interaction of the electrons as a perturbation. This method will be applied in the next section; before going to this, it is illuminating to consider the perturbation problem presented by the Eq. (163) of Section 25,

in which the electrostatic interaction of the electrons is included in the operator H and the spin variables enter only into the perturbation operator H'. The rigorous solutions of the Eq. (165) are again assumed to be known. As this equation admits the permutations as integrals, it follows that the function $u^l{}_m$, will be a member of a certain manifold of functions of $x_i y_i z_i$ (not of s_i, however) which is invariant under the permutation group. This manifold need not be one-dimensional; let its basis be $u_m{}^{lr}$, where

$$Pu_m{}^{lr} = P_{rt} u_m{}^{lt}; \tag{191}$$

the representation P_{rt} will ordinarily be irreducible; it may depend on l, but is certainly independent of m.

The zero-approximations to the solutions of Eq. (163) are linear combinations of

$$u_m{}^{lr}(x_i, y_i, z_i)\Xi_\mu{}^{sp}(\rho_i); \tag{192}$$

it is necessary to find the antisymmetric function contained in the manifold subtended by this basis.[47] Since

$$Pu_m^{lr}\Xi_\mu{}^{Sp} = P^S_{rt}P_{pq}u_m^{lt}\Xi_\mu^{Sp},$$

the character of its representation is

$$\chi(P){}^S\chi(P), \text{ where } \chi(P) = \sum P_{rr}, \text{ and } {}^S\chi(P) = \sum_q {}^S P_{qq}$$

and the manifold therefore contains

$$\frac{1}{N!}\sum_P \epsilon(P)\chi(P){}^S\chi(P)$$

antisymmetric functions. According to the last paragraph of Section 26, this sum will be zero unless P_{rt} is equivalent to the representation reciprocal to ${}^S P_{pq}$, and equal to unity when this condition is fulfilled The manifold thus contains at most one antisymmetric function, given by

$$w = \sum_P \epsilon(P) P\left\{u_m^{lr}\Xi_\mu{}^{Sp}\right\}, \tag{193a}$$

(which expression is essentially independent of the indices r and p). The number S was originally introduced as the total spin angular momentum (Section 23); the exclusion principle imposes a second role on S: it also determines the character of the representation P_{rt} based on the spin-free functions $u_m{}^{lr}$.

This result is of importance for the theory of normal multiplets: to each energy level of the spin-free equation there will be associated either zero, or else precisely $(2l+1)\ (2S+1)$ characteristic functions satisfying the Pauli exclusion principle. To the zeroth approximation these have the form

$$w_{m\mu}{}^{ls} = \sum_r\sum_P a_{rp}u_m^{lr}\Xi_\mu{}^{Sp} \tag{193b}$$

[47] Neumann and Wigner, Zeits. f. Physik 49, 80 (1928).

where the a_{rp} are independent of m and μ. These functions are a basis for the representation ${}^{l}D_{mn}{}^{s}D_{\mu\nu}$ of the group RD; the manifold subtended by them may be reduced by the method of Section 22 even though they are not single products. The basis of the reduced manifold is

$$w_M{}^J = \sum_m \sum_\mu A^{Jls}_{M m\mu} w^{ls}_{m\mu} \tag{194}$$

where $J = l+s,\ l+s-1,\ \cdots\ |l-s|$. As there is no accidental degeneracy in the inner quantum number J, it follows at once that the perturbation matrix is diagonal, so that the functions $w_M{}^J$ are themselves the zero-approximations to the exact solutions. The energy levels will depend only on J and not on M. There will thus be $2S+1$ (or $2l+1$ if $l<S$) different energy levels arising from a single energy level of the spin-free equation. These levels form a unit, the normal multiplet;[48] the number $2S+1$ is called its multiplicity.

These energy levels may be characterised by the three quantum numbers J, l, S, for according to Eq. (147), the operators L^2 and S^2 are diagonal to the approximation here considered. The selection principles

$$\Delta J = \pm 1 \text{ or } 0, \text{ but } J = 0 \rightarrow J = 0 \text{ excluded,}$$
$$\text{odd} \rightleftarrows \text{even}$$

follow immediately from the considerations of Section 24. The $w_M{}^J$ are odd or even according to the character of $u_m{}^{lr}$. These selection principles are valid to every approximation; two other approximate rules are easily deduced by noting that, because of Eq. (194), the transition amplitudes are approximately given by linear functions of

$$(X w^{lS}_{m\mu},\ w'^{l'S'}_{m'\mu'}), \text{ etc.} \tag{195}$$

Since X, Y, Z are independent of the spin variables (commute with $D(\alpha,\beta,\gamma)$) and $Dw_{m\mu}{}^{lS} = {}^{S}D_{\mu\nu} w_{m\nu}{}^{lS}$, it follows from Section 11, that each of these expressions will be zero unless

$$\Delta S = S - S' = 0.$$

Furthermore, it is known that the expressions (195) will vanish unless

$$\Delta l = \pm 1 \text{ or } 0.$$

These selection principles are identical with those deduced from the Landé theory. The quantum theory of the normal multiplet differs from the Landé theory in one important respect: namely, the dual rôle of the quantum number S. This makes itself apparent in the separations of the energy levels. According to the older theory, all levels having the same l and the same electronic configuration (cf. Section 29) should differ in energy only because of the magnetic properties of the electron. Such levels can differ in J and S, and the theory obviously demands that the energy-differences between levels of different J to be of the same order of magnitude as the

[48] Cf. F. Hund, Linien Spektren, Berlin (1927), sec. 22.

energy differences between levels of different S. It is well-known that this is not the case: in He I, for example, the levels of lines differing only in J are very closely spaced, but the energy differences between those levels of $S=0$ (singlet system) and $S=1$ (triplet system) are comparable to those between levels of different l ($nS-nP$, etc). This is readily accounted for by the foregoing analysis: even when the magnetic (spin) energies of the electrons are completely neglected, a quantum number S can be introduced which specifies the character of the representation P_{rt}. Levels having different S values will differ in energy even to this approximation (cf. Section 6). When the magnetic energy is then considered as a perturbation, the Pauli exclusion principle, Eq. (193a), requires that the total spin momentum shall be equal to this quantum number; it is thus comprehensible that the two sets of energy differences should have different magnitudes.[49]

It should be emphasized that the normal multiplet is an approximation to actual multiplets, as is obvious from the fact that only the zero approximation to the wave-function has been used in the discussion. The theory of non-normal multiplets may most readily be treated by considering both the electrostatic interaction of the electrons and their magnetic energy as perturbations.

29. **The enumeration of the terms arising from a given electronic configuration.** In Eqs. (185) and (186) the operators $H(i)$ depend on the four variables x_i, y_i, z_i, s_i. If the interaction of the spin and orbital motion be neglected, the solutions of Eq. (186) may be written

$$u_i(i) = v_i(x_i y_i z_i)\xi(\tau_i, s_i), \tag{195}$$

where $\xi(\frac{1}{2}, s_i)=\xi(s_i)$, $\xi(-\frac{1}{2}, s_i)=\eta(s_i)$. The number τ_i may be called the spin quantum number of the i-th electron, in distinction to s_i, which is the spin-variable. The N functions $v_1, v_2, \cdots v_N$ may be said to specify the electronic configuration of the atom. For the moment, it will be convenient to assume that the sequence of functions $v_1, v_2, \cdots v_N$ is given, but that the sequence of quantum numbers $\tau_1\tau_2\cdots\tau_N$ is arbitrary, subject only to the restriction that the determinant (190) must not vanish when the functions (195) are substituted into it. The first problem presenting itself is the enumeration of the sequences which satisfy this condition. As the τ's can take on only the values $\pm 1/2$, there are 2^N sequences in all, so that the number of sequences satisfying the condition cannot be greater than 2^N. For the purpose in hand, it may be assumed that the determinant will vanish only when two or more of the functions u_i are identical in form. This will certainly not be the case if all the functions v_i are different, but even if $v_i=v_j$ it may still be avoided if $\tau_i\neq\tau_j$. If more than two of the functions v_i are identical, however, then the determinant will certainly vanish.

In counting up the number of different nonvanishing determinants it will therefore be permissable to consider only the case where

[49] For an explicit treatment of this subject, see P. A. M. Dirac, Proc. Roy. Soc. **A123**, 714 (1929).

$$v_1 = v_2,\ v_3 = v_4,\ \cdots v_{2-1} = v_{2d}, \tag{196}$$

but

$$v_2 \neq v_4 \neq v_6 \cdots \neq v_{2d},$$

and

$$v_i \neq v_j$$

if either i or j is greater than $2d$. The integer d may be referred to as the number of pairs. Since τ_1 must be different from τ_2, τ_3 from τ_4, etc., it is allowed to take $\tau_1 = \frac{1}{2}$, $\tau_2 = -\frac{1}{2}$, $\tau_3 = \frac{1}{2}$, $\tau_4 = -\frac{1}{2}$, $\cdots \tau_{2d-1} = \frac{1}{2}$, $\tau_{2d} = -\frac{1}{2}$, leaving only the $N-2d$ indices $\tau_{2d+1} \cdots \tau_N$ arbitrary. For each sequence of these, the determinant will not vanish, nor will two sequences yield the same determinant. Hence there are just 2^{N-2d} antisymmetric solutions for each electronic configuration.

The next problem which presents itself is the addition of the spin vectors. The 2^{N-2d} antisymmetric functions just enumerated subtend a manifold, $\mathfrak{M}_D$, which is invariant under the group of unitary substitutions D. To see this, call the simple product of the N functions (195) $\psi(\tau_1\tau_2 \cdots \tau_N)$; then certainly

$$D\psi(\tau_1 \cdots \tau_N) = {}^{1/2}D_{\tau_1\sigma_1} \cdots {}^{1/2}D_{\tau_N\sigma_N}\psi(\sigma_1 \cdots \sigma_N). \tag{197}$$

As the operations D commute with the permutations P, and the determinant obtained from this sequence may be written

$$w(\tau_1 \cdots \tau_N) = \sum_P \epsilon(P) P\psi(\tau_1 \cdots \tau_N), \tag{198}$$

the truth of the statement follows at once. The character of the representation based on $\mathfrak{M}_D$ is readily calculable:

$$D(\omega, 0, 0)w(\tau_1 \cdots \tau_N) = e^{i\tau\omega} w(\tau_1 \cdots \tau_N) \tag{199}$$

where

$$\tau = \tau_1 + \tau_2 + \tau_3 + \cdots + \tau_N$$

and is an integer or a half-integer according as N is even or odd; therefore

$$\chi(\omega) = \sum e^{i\tau\omega}.$$

The sum is to be extended over all sequences $\tau_1 \cdots \tau_N$ which yield a non-vanishing determinant. It is most readily evaluated by first counting up the number of such sequences yielding a given value of τ; as the sum of the first 2d terms is necessarily zero, this number is

$$\binom{N-2d}{N/2-d-\tau}$$

and therefore

$$\begin{aligned} \chi(\omega) &= \sum_{\tau=-N/2+d}^{N/2-d} \binom{N-2d}{N/2-d-\tau} e^{i\tau\omega} \\ &= \sum_{S=1/2 \text{ or } 0}^{N/2-d} q(S, N-2d)^S \chi(\omega), \end{aligned} \tag{200}$$

by Eqs. (111) and (169). The possible values of the resultant spin are therefore

$$S = N/2 - d,\ N/2 - d - 1,\ \cdots\ (\tfrac{1}{2} \text{ or } 0).$$

Values of S greater than $N/2-d$ are thus excluded by the Pauli principle, as was evident from its physical statement. These considerations have led to a further conclusion, however: the representation S does not occur $q(S, N)$ times in the manifold of antisymmetric functions, but only $q(S, N-2d)$ times. The Pauli principle has also reduced the number of terms having a given value of S below that to be expected from the considerations of Section 25. This might also have been deduced from the vector model. The actual reduction of the basis (198) to the form w_μ^{Sp}, where $p=1, 2, \cdots q(S, N-2d)$, may readily be accomplished by the method already described.

It remains to consider the effect of the rotation group on the antisymmetric functions of $\mathfrak{M}_D$. This manifold is not invariant under the operations R, a circumstance which arises from a somewhat too narrow interpretation of the term "electronic configuration." The functions v_i of Eq. (195) are each characterized by three indices n, l, m, the principal, azimuthal, and magnetic quantum numbers. In assuming the sequence of functions $v_1 \cdots v_N$ to be given, it is assumed that all three indices of each function are specified. In the spectroscopic sense of the term "electronic configuration," only the n and l of each v_i is specified, the m being left arbitrary. This is the sense in which the term will be used hereafter: to each electronic configuration there will correspond a number of manifolds like $\mathfrak{M}_D$, one based on each of the possible sequences of functions v_i which can be obtained by different assignments of the magnetic quantum numbers. These possible sequences are restricted in two ways: the conditions (196) must be fulfilled, and $|m| \leqq l$ for each v. The antisymmetric functions obtained from each of these sequences will belong to the same zero-approximation energy level. The various manifolds $\mathfrak{M}_D$ will not all be identical, but may differ in the number of pairs in the sequence $v_1 \cdots v_N$ on which they are based. The manifold containing all of them will be invariant under R, as may be shown by a slight modification of the reasoning of Eqs. (197) and (198). This manifold is already partially reduced: the manifolds $\mathfrak{M}_\mu^s$ which contain all functions having a given S and μ are themselves invariant under the group R, since $RD=DR$. The $(2S+1)$ manifolds having a given S all have the same structure, so that it is necessary to consider the reduction of only one of them.

These considerations are very complex if general cases are considered, but become simple for any special case.[50] For this reason, only two examples will be treated: the first is that of an atom containing two electrons each having $n=2$ $l=1$. In spectroscopic notation, the configuration is p^2. The foregoing considerations are summarized in Table II.

[50] Cf. Wigner, Zeits. f. Physik **43**, 651 (1928) for other examples.

TABLE II. *Configuration* p^2.

m_1	m_2	d	$q(S, N-2d)$		m
			$S=1$	$S=0$	
1	0	0	1	1	1
1	−1				0
0	−1				−1
1	1	1	0	1	2
0	0				0
−1	−1				−2

The first two columns contain the six pairs of values m_1, m_2 which result in essentially different determinants. The second column contains d, the number of pairs. Each of these sequences results in a manifold $\mathfrak{M}_D$ containing 2^{2-2d} functions; these are reducible into submanifolds containing $(2S+1)$ $q(S, 2-2d)$ functions each, as is shown in the fifth and sixth columns. The submanifolds $\mathfrak{M}_\mu^S$ which are invariant under the rotation group are four in number: $\mathfrak{M}_1^1$, $\mathfrak{M}_0^1$, $\mathfrak{M}_{-1}^1$, and $\mathfrak{M}_0^0$, corresponding to the two values 1 and 0 of S. The characters of these are easily calculated: the action of $R(\omega, 0, 0)$ on any of the functions is to multiply it by $e^{im\omega}$, where m is the sum of the entries in the first two columns of the row corresponding to the function. The manifold $\mathfrak{M}_1^1$, for example, contains three functions for which $m=1$, 0, -1, respectively; its character is therefore ${}^1\chi$. The same is true of $\mathfrak{M}_0^1$ and $\mathfrak{M}_{-1}^1$. Hence each of these manifolds is irreducible under the group R and the nine functions contained in them may be designated by $w_{m\mu}^{11}$, having the properties

$$Rw_{m\mu}^{11} = {}^1D_{mn}w_{n\mu}^{11},$$
$$Dw_{m\mu}^{11} = {}^1D_{\mu\nu}w_{m\nu}^{11}.$$

The manifold $\mathfrak{M}_0^0$ is slightly more complex; it contains six functions, and its character is

$$\chi(\omega) = e^{2i\omega} + e^{i\omega} + 2 + e^{-i\omega} + e^{-2i\omega}$$
$$= {}^2\chi(\omega) + {}^0\chi(\omega).$$

It is therefore reducible into one manifold $L=2$ and one $L=0$. Its functions are thus w_{m0}^{20} and w_{00}^{00} for which

$$Rw_{m0}^{20} = {}^2D_{mn}w_{n0}^{20},$$
$$Dw_{m0}^{20} = w_{m0}^{20};$$

and

$$Rw_{00}^{00} = w_{00}^{00},$$
$$Dw_{00}^{00} = w_{00}^{00}.$$

These various functions will be linear combinations of the determinants calculated according to the procedure outlined in the earlier part of this

section. The coefficients of these combinations are again the A's of Section 22. On combining the S and L vectors, it is seen that the final result is one ^3P, one ^1D and one ^1S multiplet. All of these arise from the electronic configuration p^2.

As a second example consider the configuration p^3s. The table (Table III) is arranged in the same manner as before. In connection with the first row,

TABLE III. *Configuration p^3s.*

$l=1$			$l=0$	d	2^{N-2d}	$q(N-2d)$			m
m_1	m_2	m_3	m_4			$S=2$	$S=1$	$S=0$	
1	0	−1	0	0	16	1	3	2	0
1	1	0	0						2
1	1	−1	0						1
1	0	0	0	1	4	0	1	1	1
−1	0	0	0						−1
1	−1	−1	0						−1
0	−1	−1	0						−2

a convenient check against mistakes in the preparation of such a table may be noted: the total number of functions in the $\mathfrak{M}_D$ corresponding to this row is 16, and this must be equal to

$$\sum_S (2S+1)q(S_1N-2d) = 5\cdot 1 + 3\cdot 3 + 1\cdot 2 = 16.$$

The functions enumerated in this table fall into nine manifolds invariant under R: $\mathfrak{M}_\mu{}^2$, $\mathfrak{M}_\mu{}^1$, and $\mathfrak{M}_0{}^0$. The character of $\mathfrak{M}_\mu{}^2$ is primitive: $^0\chi$, but the characters of the others are reducible:

$$\mathfrak{M}_\mu{}^1\colon \chi(\omega) = e^{2i\omega} + 2e^{i\omega} + 3 + 2e^{-i\omega} + e^{-2i\omega} = {}^2\chi(\omega) + {}^1\chi(\omega) + {}^0\chi(\omega),$$

$$\mathfrak{M}_0{}^0\colon \chi(\omega) = e^{2i\omega} + 2e^{i\omega} + 2 + 2e^{-i\omega} + e^{-2i\omega} = {}^2\chi(\omega) + {}^1\chi(\omega).$$

Hence the forty[51] antisymmetric functions listed can be combined into linear combinations $w_{m\mu}{}^{LS}$ with the following values of L and S:

L	0	2	1	0	2	1
S	2	1	1	1	0	0

The coefficients of these linear functions can be calculated by a simple extension of the methods explained in Sections 22 and 25. On adding the L and S vectors, it is obvious that one multiplet of each of the types ^5S, ^3D, ^3P, ^3S, ^1D, and ^1P will result.

One other case may be considered; that of an atom containing $2(2l+1)$ electrons of the same n and l. In this case the sequence of m-values is determined uniquely, for there are just $2l+1$ different values available and none may be used more than twice, hence each must be used exactly twice. Their

[51] It may not be out of place to remark that the examples are chosen from among the more complex, rather than from the simplest.

sum must then be zero, and since $N-2d=0$, $S=0$, and $q(0, 0)=1$, there is just one manifold, $L=S=J=0$. This is the explanation of the phenomenon of "closed shells."

It is obvious from the foregoing examples that the complete manifold of zero approximations for any unperturbed energy level will usually contain several irreducible submanifolds having the same character. In the first example, there are two manifolds for $J=2$, and two for $J=1$. These differ in the values of L and S from which they arise. The perturbation matrix will thus not be diagonal, but have the form $H_{\alpha\beta}(J)\,\delta(J, J')\,\delta(M, M')$ where α, β are indices referring to functions having the same J and M (cf. Section 6). The zero approximations to the characteristic functions will therefore differ from those considered at the end of Section 26 in that their terms do not all arise from the same L and S. The multiplets enumerated in the foregoing examples will not be "normal," and the two approximate selection principles $\Delta S=0$, $\Delta L=0, \pm 1$ will not be valid. This is an indication of the fact that this method of calculation is capable of yielding a better approximation than the former method.

30. **Evaluation of the perturbation matrix.** In actually carrying out the perturbation calculation outlined in the foregoing section, it is first necessary to evaluate inner products of the form

$$(Hw, w') \tag{201}$$

where w and w' are two N-rowed determinants and $PH=HP$. The evaluation of these is not difficult in the cases here under consideration. The determinants may conveniently be written in the form

$$w = \frac{1}{(N!)^{1/2}} \sum_P \epsilon(P) P\psi \tag{202}$$

$$w' = \frac{1}{(N)^{1/2}} \sum_P \epsilon(P) P\phi,$$

in which

$$\begin{aligned} \psi &= u_1(1)u_2(2)\cdots u_N(N) \\ \phi &= v_1(1)v_2(2)\cdots v_N(N) \end{aligned} \tag{203}$$

and $01/(N!)^{1/2}$ is a normalizing factor. The functions $u_i(i), v_i(i)$ depend on all four variables x_i, y_i, z_i, s_i, and all the u's are different from one another, as are all the v's.

Substituting (202) into (201),

$$\begin{aligned} (Hw, w') &= \frac{1}{N!} \sum_P \sum_Q \epsilon(P)\epsilon(O)(PH\psi, Q\phi) \\ &= \frac{1}{N!} \sum_P \sum_Q \epsilon(P)\epsilon(Q)(PHQ^{-1}\psi, \phi) \end{aligned}$$

since the permutations are orthogonal; since they commute with H, this is

$$= \frac{1}{N!} \sum_P \sum_Q \epsilon(PQ^{-1})(PQ^{-1}H\psi, \phi).$$

Applying the rearrangement theorem to $\sum_P$, this becomes

$$\begin{aligned}(Hw, w') &= \frac{1}{N!} \sum_P \sum_Q \epsilon(P)(PH\psi, \phi) \\ &= \sum_P \epsilon(P)(PH\psi, \phi),\end{aligned} \tag{204}$$

which contains at most $N!$ terms, as compared with the $(N!)^2$ terms initially to be considered. In actual cases, many of these remaining terms are zero; this arises from the fact that H always has the form

$$H = \sum_{i=1}^{N} H(i) \tag{205}$$

where $H(i)$ depends only on the four variables i, or else the form

$$H = \sum_{(i,j)} H(i, j) \tag{206}$$

where $H(i, j)$ depends only on the eight variables i, j.

In the first case, one is led to consider inner products of the type

$$(PH(i)\psi, \phi). \tag{207}$$

The functions entering into it will have the form

$$P\left\{H(i)u_i(i) \prod_{k \neq i} u_k(k)\right\}$$

and

$$v_i(i) \prod_{k \neq i} v_k(k).$$

If the inner product were merely a $3N$-fold integral (Eq. (3)), the expression (207) would be expressible as a product of N three-fold integrals; when it has the form of Eq. (3a), it can still be expressed as the product of N inner products each involving only four variables x_k, y_k, z_k, s_k, and defined by

$$[f(k), g(k)]_k = \sum_{s=-1/2}^{+1/2} \int\int\int f(x_k y_k z_k s) g^*(x_k y_k z_k s) dx_k dy_k dz_k.$$

Furthermore, the sequences $u_k(k)$, $v_k(k)$ will, in most actual cases, satisfy the following relations:

$$\begin{aligned}[u_k(1), u_j(1)]_1 &= \delta_{kj}, \\ [v_k(1), v_j(1)]_1 &= \delta_{kj}, \\ [u_k(1), v_j(1)]_1 &= c_k \delta_{kj}.\end{aligned} \tag{208}$$

In the last equation, c_k is either zero or one; this follows from the fact that the sequences $u_1 \cdots u_N$ and $v_1 \cdots v_N$ will usually differ only by a relatively few functions which are orthogonal to one another. Because of Eq. (208), it is seen that the expression (207) will certainly vanish unless $P = I$, and in that case, reduce to

$$(H(i)\psi, \phi) = [H(i)u_i(i), v_i(i)]_i \prod_{k\neq i} c_k.$$

Even this will vanish unless all the c_k, except possibly c_i, are different from zero. When this condition is fulfilled, with $c_1 = 0$

$$(Hw, w') = [H(1)u_1(1)v_1(1)]_1; \tag{209a}$$

in the particular case $w' = w$; however, none of the c's vanish and

$$(Hw, w) = \sum_{i=1}^{N} [H(i)u_i(i), u_i(i)]_i. \tag{209b}$$

When H has the form (206) inner products of the type

$$(PH(ij)\psi, \phi) \tag{210}$$

must be considered. These will certainly vanish unless P is either the identity or the (odd) permutation which interchanges i and j and leaves all other variables unchanged. In these particular cases, the expressions reduce to

$$[H(ij)u_i(i)u_j(j), v_i(i)v_j(j)]_{ij} \prod_{k\neq i,j} c_k$$

and

$$[H(ij)u_i(j)u_j(i), v_i(i)v_j(j)]_{ij} \sum_{k\neq i,j} c_k$$

where $[\ \ ]_{ij}$ is an inner product involving only the eight variables i and j. These still vanish unless all c_k, except possibly c_i and c_j, are equal to 1. There are now three cases to distinguish:

1. $c_1 = c_2 = 0$, all other $c_k = 1$; then

$$\begin{aligned}(Hw, w') = {} & [H(12)u_1(1)u_2(2), v_1(1)v_2(2)]_{12} \\ & - [H(12)u_1(2)u_2(1), v_1(1)v_2(2)]_{12};\end{aligned} \tag{211a}$$

2. $c_1 = 0$, all other $c_k = 1$;

$$\begin{aligned}(Hw, w') = {} & \sum_{i=2}^{N} \{ [H(1, i)u_1(1)u_i(i), v_1(1)v_i(i)]_{1i} \\ & - [H(1, i)u_1(i)u_i(1), v_1(1)v_i(i)]_{1i} \};\end{aligned} \tag{211b}$$

3. $w = w'$, all $c_k = 1$;

$$\begin{aligned}(Hw, w') = {} & \sum_{(i,j)} \{ [H(ij)u_i(i)u_j(j), u_i(i)u_j(j)]_{ij} \\ & - [H(ij)u_i(j)u_j(i), u_i(i)u_j(j)]_{ij} \}.\end{aligned} \tag{211c}$$

The calculation of the perturbation matrix in terms of the determinants w which formed the starting point of the last section is thus reduced to a relatively simple form. By carrying out the canonic transformations involved in the reduction of the manifold of antisymmetric functions, this matrix can be reduced to the form of Eq. (47). The system of equations thus obtained is in its simplest form and its solution cannot be avoided if the characteristic functions are to be determined. Its characteristic values may often be determined by the method of sums without actual solution of the secular equation, as has been explained.

VI. The Influence of External Fields

31. **The Zeeman effect. The sum rules for the g-factors.** The energy levels of an atom when not acted on by an external force are, as has been seen, always degenerate. The manifold of wave functions corresponding to each is $2J+1$ dimensional[52] and its basis may be supposed given as u_M^J. Each of these functions is the basis of one of the irreducible representations of the group S, K', $K'S$, which dominates the problem when an homogeneous magnetic field $\mathfrak{H}$ is applied to the atom.

The perturbation operator which must be added to the operator of the wave equation in the absence of the field is

$$\frac{eh\mathfrak{H}}{4\pi mc}(L_z + 2S_z) = \frac{eh\mathfrak{H}}{4\pi mc}(J_z + S_z) \tag{212}$$

where J_z, L_z, S_z are the z-components of the operators associated to the total, orbital, and spin angular momenta, respectively. The perturbation matrix arising from this operator must be diagonal, since there is no accidental degeneracy and the representation of the manifold is completely reduced. This result may be made more precise by noting that J_z+S_z is the z-component of a vector operator, and that therefore, according to the results of Section 24,

$$(J_z + S_z)u_M^J = gMu_M^J \tag{213}$$

where g is a factor independent of M, but dependent on J. The action of the magnetic field will therefore be to split up the degenerate energy level into $2J+1$ single levels, which are equally spaced.

The approximate values of the g-factors can be determined in the case of normal multiplets, and partially determined in the more general cases. In the former case, the u_M^J are approximately of the form

$$u_M^J = \sum_m \sum_\mu A^{JLS\ LS}_{M m\mu} v_{m\mu}, \tag{214}$$

so that

$$g(J) - 1]Mu_M^J = S_z u_M^J = \sum_m \sum_\mu A^{JLS\ LS}_{M m\mu} v_{m\mu}. \tag{215}$$

This equation may be solved for $g(J)$ in several ways, the simplest depending on Eq. (148).[53] The method which will be used here is analogous to the method of sums, and is preferred because it may be extended to the case of a general multiplet. In Eq. (214) J takes on the values $L+S$, $L+S-1$, $\cdots |L-S|$, and $|M| \leqq J$; the equations may also be written, according to Section 22:

$$u^{L+S}_{L+S} = v^{LS}_{LS};$$

$$u^{L+S}_{L+S-1} = a v^{LS}_{(L-1)S} + b v^{LS}_{L(S-1)},$$

$$u^{L+S-1}_{L+S-1} = c v^{LS}_{(L-1)S} + d v^{LS}_{L(S-1)}; \quad \text{etc.}$$

[52] For the special case $J=0$, this degeneracy disappears, but it is not necessary to treat this case separately.

[53] Weyl, *l.c.* p. 164.

From the first of these and Eq. (215), it follows that

$$(L+S)[g(L+S)-1] = S$$

or

$$g(L+S)-1 = S/(L+S). \tag{216a}$$

From the second, using the fact that the diagonal sum of the matrix associated to S must be invariant under the canonic transformation,

$$(L+S-1)\{[g(L+S)-1]+[g(L+S-1)-1]\} = 2s-1$$

or

$$g(L+S-1)-1 = (2S-1)/(L+S-1) - S/(L+S). \tag{216b}$$

The three equations which follow those written out above determine u^{L+S}_{L+S-2}, u^{L+S-1}_{L+S-2}, and u^{L+S-2}_{L+S-2} in terms of $v^{LS}_{L(S-2)}$ $v^{LS}_{(L-1)(S-1)}$, and $v^{LS}_{L(S-2)}$. Reasoning in the same manner, it is found that

$$g(L+S-2)-1 = (3S-3)/(L+S-2) - (2S-1)/(L+S-1). \tag{216c}$$

Continuing, it is readily found by induction, that

$$g(J)-1 = \frac{J(J+1)-L(L+1)+S(S+1)}{2J(J+1)} \tag{216}$$

which is Landé's formula.

This method is capable of yielding some information regarding the approximate values of $g(J)$ in the case of the general multiplets. As an example, let it be required to discuss the g-factors of the levels arising from the configuration *sd*. In this case, the considerations of Section 29 indicate that two values of S, namely 0 and 1, are associated to each energy level when the mutual interaction is not considered. The basis of the manifold of this level may be taken to be v^{20}_{m0}, $v^{21}_{m\mu}$. The effect of the mutual interaction is then, approximately, merely to reduce this basis to ${u_M}^3$, ${u'_M}^2$, ${u''_M}^2$ and ${u_M}^1$, these functions being linear combinations of the original v's. The characteristic to be noted is that there are two sets of u's for which $J=2$; if the coupling were "normal," these functions would correspond to a ^3D and a ^1D normal multiplet. The same procedure is to be followed as before:

$${u_3}^3 = v^{21}_{21},$$

hence

$$3[g(3)-1] = 1, \text{ or } g(3) = 4/3.$$

The next step encounters difficulties, however:

$$\begin{aligned} {u_2}^3 &= a v^{20}_{20} + b v^{21}_{20} + c v^{21}_{11} \\ {u'_2}^2 &= a' v^{20}_{20} + b' v^{21}_{20} + c' v^{21}_{11} \\ {u''_2}^2 &= a'' v^{20}_{20} + b'' v^{21}_{20} + c'' v^{21}_{11} \end{aligned}$$

and hence there is only one equation for two unknowns:

$$2\{[g(3)-1]+[g'(2)-1]+[g''(2)-1]\}=0+0+1$$

whence all that can be deduced is that

$$g'(2)+g''(2)=13/6.$$

For $g(1)$, the result is determinate again:

$$g(1)=\tfrac{1}{2}.$$

On comparing these results with the g-factors for a normal ^3D, viz., $g(3)=4/3$, $g(2)=7/6$, $g(1)=1/2$, and a normal ^1D, $g(2)=1$, the general rule becomes obvious: If several levels having the same J-value arise from the same electronic configuration, the sum of their g-factors is equal to the sum of the g-factors calculated from Eq. (216) for the normal multiplets. It may also be remarked that this rule is independent of the field strength and thus governs the Paschen-Back effect.

32. **Stark effect.**[54] The manifold of functions associated to each energy level in the absence of an external field when specified in terms of a basis u_{-M}^{J} is also completely reduced to its submanifolds which are invariant under the group S, K'', SK'' which governs the problem when an homogeneous electric field is applied to the atom. The basis of the irreducible representation of this group spcified by the number M is

$$u_M{}^J,\ u_{-M}{}^J$$

or $u_0{}^J$ if J is integral and $M=0$. The action of the electric field, if J is integral, is thus to split up the energy level into J levels each two-fold degenerate, and one non-degenerate level. If J is half integral, there will be $J+1/2$ two-fold degenerate levels.

The perturbation energy in this case is

$$-eEZ,$$

E being the strength of the field and Z the sum of the rectangular coordinates of the electrons in the direction of E. Since

$$(Zu_M{}^J,\ u_M{}^J),$$

and

$$(Zu_M{}^J,\ u_{-M}{}^J),$$

are both zero (Z is an odd function) the perturbation matrix vanishes identically in this case. Hence the splitting of the energy level will be zero to the first approximation[55] The separation between the different levels will be proportional to E^2. There will thus be no linear Stark-effect unless there is accidental degeneracy—i.e., unless two levels of different J have the same energy value. This is the case only in one actual example—that of hydrogen.

[54] E. Wigner, Zeits. f. Physik **43**, 643 (1928).
[55] Cf. Neumann and Wigner, l.c.

Volume 2 *October, 1930* *Number 4*

REVIEWS
OF
MODERN PHYSICS

THERMIONIC EMISSION

By Saul Dushman
Research Laboratory of the General Electric Company

TABLE OF CONTENTS

Introductory Remarks 382
I. Technique of Electron Emission 383
A. Equations for emission, definition of work function 383
B. Remarks on vacuum technique 384
C. Forms of hot cathode tubes 385
D. Temperature scales 387
E. Correction for lead-losses 390
F. Corrections for effect of anode voltage (Schottky theory) Experimental tests of theory 392
II. Electron Emission Data for Elementary Substances 394
Table of emission constants 394
Emission constants for platinum 396
Emission constants for zirconium and hafnium 397
III. Electron Emission from Monatomic Films 398
A. *Thorium on tungsten* 398
Thermal treatment, rates of production, diffusion and evaporation 398
Variation in emission constants with θ 400
Emission at $T=1900$ as function of θ 401
Emission, efficiency of emission, diffusion constant and rate of evaporation (for completely covered film) as function of temperature 402
B. *Films of other rare earth metals and barium on tungsten and molybdenum* 403
Emission constants, rates of diffusion and of evaporation 403
C. *Caesium on tungsten and oxidized tungsten* 405
Emission as function of pressure of caesium and of temperature of filaments 405
Electron emission constants 407
Emission of positive ions 410
Rate of activation 411

Equilibrium values of θ as function of temperature, work function for electrons and positive ions..... 413
IV. Calorimetric Determination of Work Function for Metals..... 414
Table of values..... 415
V. Electron Emission from Metals at the Melting Point..... 416
Emission constants for copper..... 416
Emission constants for silver and gold..... 417
VI. Oxide Coated Cathodes..... 418
A. Preparation of cathodes..... 418
B. Temperature scales..... 421
C. Emission constants..... 423
Table of values..... 423
Characteristics of Western Electric filaments..... 427
Values of emission at $T=1000$..... 427
D. Calorimetric determinations of work function..... 430
E. Activation and emission phenomena..... 431
Thermal and electrolytic theories of formation..... 432
Change in emission constants during activation..... 436
Evaporation and diffusion of barium and oxygen..... 438
Effect of oxygen on activity of cathode..... 440
Theory of transfer of electrons in coating..... 441
VII. Effect of Gases and Positive Ion Bombardment on Electron Emission..... 444
Effect of gases on W, Mo, Ta, Pt, etc..... 445
Sputtering of thorium by positive ions..... 449
VIII. Photo-Electric Threshold and Work Function..... 451
IX. Contact Difference of Potential..... 454
Measurements on contact potentials..... 455
X. Thermionic Constant, A..... 455
A. Observations on relation between log A and b..... 456
B. Derivation of A on basis of thermodynamics..... 457
C. Derivation of A on basis of new electron theory..... 461
D. Application of wave mechanics..... 466
XI. Electron Emission in Intense Electric Fields..... 470
Experimental results and interpretation on basis of wave mechanics..... 470
XII. Thermal Emission of Positive Ions..... 473
A. Emission due to vapors of alkali metals..... 473
B. Emission from catalysts..... 475
C. Emission of positive ions of refractory metals..... 476

Introductory Remarks

THERMIONIC devices have become such an essential factor in so many purely scientific and technical applications that it is difficult to realize at the present time that only sixteen years ago the very existence of a pure electron emission from incandescent solids was questioned by many physi-

cists of good repute.[1] The apparently trivial observations made by Edison on the discharge of negative electricity from the carbon filament of an incandescent lamp to an auxiliary electrode in the bulb was the beginning of a series of scientific investigations carried out by O. W. Richardson, A. Wehnelt, I. Langmuir, W. Schottky and a large number of other investigators. Their observations have led to the development of a number of hot cathode devices, the application of which in the radio and electrical industry in general has been of far reaching importance.

The use of the hot cathode is becoming increasingly important not only in these high vacuum devices, but also in connection with various types of gas discharges and arcs. It has appeared, therefore, that it might be well to take stock, as it were, of the state of our knowledge of thermionic phenomena at the present time. There is a certain measure of reason for this as a considerable period has intervened since the publication of the classical work on the subject by O. W. Richardson,[2] while the most recent summaries are available only in German treatises, such as those of W. Schottky.[3] To these excellent discussions of the subject the writer wishes here to express his indebtedness in the preparation of the following paper.

I. Technique of Electron Emission

A. Equations for emission. On the basis of certain theoretical considerations, which will be discussed in a subsequent section, O. W. Richardson derived an equation for electron emission as a function of the temperature of the form

$$I = aT^{1/2}\epsilon^{-b/T} \tag{1}$$

where I = emission per unit area,

T = absolute temperature,

and a and b are constants characteristic of the emitting surface.

Although Richardson[4] and M. v. Laue[5] had also pointed out that an equation of the form

$$I = AT^{2}\epsilon^{-b_0/T} \tag{2}$$

would be just as valid theoretically, and would be in satisfactory agreement with observed data, it was shown by S. Dushman[6] that on the basis of the third law of thermodynamics, A in this equation should be a universal constant having the value 60.2 amps/cm^2 per deg^2 while b_0 should vary with the nature of the emitter.

As will be discussed more fully in subsequent sections, b or b_0 is a measure

[1] In this connection it is of historical interest to read the discussion on this point by I. Langmuir in Proc. Inst. Rad. Eng. **3,** 261 (1915), and G. E. Review **18,** 327 (1915).

[2] O. W. Richardson, The Emission of Electricity from Hot Bodies, Longmans, Green and Company (1921).

[3] W. Schottky, H. Rothe and H. Simon, Wien-Harms' Handbuch der Experimentalphysik **13,** Akadem. Verlagsgesellschaft, m.b.H. Leipzig (1928).

[4] O. W. Richardson, Phil. Mag. **28,** 633 (1914).

[5] M. v. Laue, Jahrb. d. Elektronik u Radioakt. **15,** 205, 257 (1918).

[6] S. Dushman, Phys. Rev. **21,** 623 (1923).

of the latent heat of evaporation of the electrons, i.e., the energy necessary to get the electrons through the surface. While these are expressed in degrees Kelvin in the above equations, it is also customary to express them in terms of volts by means of the relation

$$b_0 k = \Phi_0 e \tag{3a}$$

where k denotes Boltzmann's constant, e, the charge on the electron, and Φ_0 is known as the "work function." Substituting the well-known values for e and k, it is readily shown that

$$\Phi_0 = 8.62 \times 10^{-5} b_0 \text{ (volts)}. \tag{3b}$$

It has been the object of a large number of investigators to determine the values of these emission constants a and b(or A and b_0) as accurately as possible, and in the following sections the data thus obtained will be reviewed. Since it has been found actually that the observed value of A in equation (2) is not independent of the composition of the emitting surface, a great many theoretical investigations have been published with the object of explaining this apparent anomaly. While these views will be discussed in a subsequent section, it would seem at the present time that the reliable data available are insufficient to draw any definite conclusions on this point. Therefore more emphasis should be laid on experimental investigations to obtain accurate emission data on those substances for which data are at present unavailable, and furthermore it is necessary to obtain data, independent of emission measurements, on the values of the work functions themselves, since from such data combined with observations on emission more exact values of the constant A could be obtained.

B. General remarks on conditions requisite for accurate determinations of emission constants. From equation (2) it is seen that

$$\frac{dI/I}{dT/T} = \frac{b_0}{T} + 2.$$

For tungsten, $b_0 = 52{,}400$, and therefore for $T = 2400°\text{K}$, the temperature coefficient of emission is approximately 24. In general, the value of b_0 is lower the lower the range of temperatures at which emission data can be obtained under practical conditions, so that the ratio b_0/T does not vary within very wide limits.

Thus the observations on emission should be made under conditions which will make it possible to determine the temperature accurately. Furthermore, since the emitting source is either in the form of a filament, or an extended surface, care has to be taken to determine as accurately as possible the actual area at the maximum temperature, and to introduce proper corrections, where necessary, for temperature gradients along the emitting surface. The most important precautions, however, in connection with observations on emission are those regarding the removal of adsorbed and occluded gases from the emitter itself, and the necessity of maintaining as

nearly perfect vacuum conditions as possible. Most of the observations on emission made up to 1914, and a considerable number of those made since then, are almost worthless because of the poor vacuum conditions under which they were made. As a result largely of the work of I. Langmuir and his associates it has become recognized that adsorbed gases may alter the emission profoundly, and, furthermore, that if there is sufficient residual gas in the hot cathode tube, positive ion bombardment tends to decrease the emission, especially from coated surfaces or those covered with monatomic films of "active" substances.

At the present time it is not necessary to describe in detail the methods available for obtaining ideal vacuum conditions and for removing adsorbed gases from the surface of the emitter. These are described in several treatises on this subject,[7] as well as in numerous papers. It is also hardly necessary to remark that all stop-cocks, greased joints, etc., should be avoided in connection with the exhaust and preparation of tubes containing cathodes for which it is desired to determine the electron emissivity.

While the evaporation of tungsten with the bulb immersed in liquid air was used by Langmuir in his first investigations in this field, other "getters" (clean-up agents for producing high vacuum) have come into use with the advent of vacuum devices for radio. For this purpose, magnesium, calcium, barium and alloys of rare-earth metals ("Mischmetal") have been utilized. According to investigations carried out in this laboratory by Mrs. M. R. Andrews, barium cleans up practically all residual gases at ordinary temperatures, while magnesium is ineffective in the case of hydrogen, and calcium does not take up nitrogen to any great extent.

Extremely low pressures may also be obtained by the use of a side tube containing charcoal (which has previously been well exhausted) immersed in liquid air. In this case, however, care has to be taken to see that the liquid air is maintained at constant level during the series of measurements.

C. Forms of hot cathode tubes for investigation of emission. One of the simplest forms of hot cathode devices is that used by I. Langmuir[8] in his first investigations on the emission from tungsten and thoriated tungsten. Two filaments, one of tungsten, and the other of the metal under investigation are sealed into a lamp bulb. Platinum flush seals are inserted in the walls of the bulb opposite to the filaments, and the tube is exhausted on a high vacuum system consisting of mercury condensation pump with liquid air trap between bulb and pump. During exhaust, the bulb is baked out

[7] *References on High Vacuum Technique*

(1) L. Dunoyer, Vacuum Practice, G. Bell and Sons, London, 1926.
(2) F. H. Newman, The Production and Measurement of Low Pressures, D. Van Nostrand Company, New York, 1925.
(3) G. W. C. Kaye, High Vacua, Longmans, Green and Company, 1927.
(4) S. Dushman, High Vacuum, G. E. Review, 1922, also translated by R. G. Berthold and E. Reimann, Julius Springer, Berlin (1926).
(5) A. Goetz, Physik u. Technik des Hochvakuumes, Friedr. Vieweg and Sohn, Akt. Ges., Braunschweig (1926).

[8] I. Langmuir, Phys. Rev. **2**, 450 (1913); also Phys. Zeits. **15**, 516 (1914).

for about one hour to the maximum temperature to which the particular glass may be subjected without collapsing. Then the filaments are flashed for a few minutes at very high temperatures (just below the melting point), and for a longer time at lower temperatures (2400°K approximately for tungsten), to eliminate occluded gases and to dissociate or evaporate any oxide films on the surface. This high temperature treatment should be continued until, as indicated by a proper type of low pressure gauge, the evolution of gas has practically ceased, and the pressure in the tube is as low as possible. The bulb is sealed off, immersed in liquid air, and the tungsten filament heated at the same time to a temperature of about 2900–3000°K. At these temperatures the tungsten evaporates fairly rapidly, "cleans up" residual gases, and forms a deposit on the walls which is subsequently used as anode in the emission observations. The evaporation should be continued until the resistance of the deposit as measured between the flush seals is less than 100 ohms.

The tungsten filament and deposit are used as anode in the subsequent emission measurements on the second filament, and it is advisable to carry out these measurements with the bulb immersed in liquid air in order to clean up any further traces of gases which may be evolved from the heated cathode.

The filaments used should preferably have comparatively low voltage drop, but should be long enough to minimize the effect of errors in correcting for lead losses (see below.) To determine the temperature of the filament, accurate observations should be made on the voltage drop and filament current.

A more convenient modification of Langmuir's arrangement has been used by S. Dushman[9] and his associates in precision measurements on electron emission from various metals. Instead of volatilizing tungsten on the walls, a piece of calcium wire inserted in a tungsten spiral is used and the metal evaporated while the tube is on the pump. Care should be taken in using this method to see that the calcium wire used is not badly oxidized and the tube should be exhausted as soon as possible after the wire has been sealed in.

K. H. Kingdon[10a] and K. H. Kingdon and I. Langmuir[10b] have used a form of tube which, while more complicated in construction, has the advantage that no corrections have to be made for lead losses. The filament under investigation is suspended along the axis of three co-axial cylinders of similar diameter and arranged with minimum distance of separation between them. The emission is measured to the central cylinder, while the two end cylinders are connected to the filament and act as "guard rings" to prevent any electrons emitted from the ends of the filament from reaching the central anode.

When the material for which it is desired to obtain emission data cannot be drawn in wire form and is available only in the form of sheets, both the

[9] S. Dushman *et al*, Phys. Rev. **25**, 338 (1925): **29**, 857 (1927).
[10a] K. H. Kingdon, Phys. Rev. **24**, 510 (1924).
[10b] K. H. Kingdon and I. Langmuir, Phys. Rev. **22**, 148 (1923).

problem of heating the surface to a uniform temperature and that of measuring the maximum temperature accurately becomes very difficult. In the case of molten metals, A. Goetz has overcome these difficulties by an ingenious method which will be described in connection with the discussion of his observations.

D. Temperature scale. For a number of substances data are now available by means of which very accurate determinations may be made of the temperature of the emitting surface. The results of the large number of investigations on this subject have been reviewed comprehensively by E. Lax and M. Pirani.[11] Data on the total radiation from a black body an various solids have also been summarized by W. W. Coblentz.[12] References to some of the more important materials are given in the following section.

Undoubtedly the most accurate method for determining the temperature consists in the measurement of the brilliancy (candles per cm^2). The temperature coefficient, that is, the ratio of dB/B to dT/T (where B denotes the brightness in international candles per cm^2) for tungsten varies from 22.75 at $T=1000°K$ to 8.45 at $T=3000°K$.[13] Thus with the exception of the electron emission and the rate of evaporation, the candle power shows the greatest variation with temperature, and therefore any inaccuracy in the determination of this value involves less error in that of the temperature than is involved in the determination of any other property of the material, such as resistance or watts radiated.

While the values of B as a function of T have been determined for a number of substances, it is possible to obtain a fair approximation to the true temperature of a material for which the average luminous emissivity[14] is unknown by the following methods.

By means of a photometer it is possible to determine the temperature T_c, at which the material emits light of the *same color* as a black body. This gives a value for the temperature which is higher than the true temperature. For tungsten the color temperatures corresponding to different values of the true temperature as reported by Worthing and Forsythe are given in Table I.

TABLE I. *Relation between color temperature (T_c), brightness temperature (T_s) and true temperature (T) for tungsten.*

T_c for tungsten	T	$(T_c-T)/T$	T_s for tungsten
1006	1000	0.006	966
1517	1500	.011	1420
2033	2000	.0165	1857
2557	2500	.023	2274
3094	3000	.031	2673

[11] Handbuch der Physik **19,** 1–45, **21,** 190–272, Julius Springer, Berlin, (1929).

[12] International Critical Tables **5,** 238–245 (1929).

[13] H. A. Jones and I. Langmuir, G. E. Review **30,** 310, 354, 408 (1927).

[14] By average luminous emissivity is meant the ratio of the total normal brightness to that of a black body at the same temperature. For tungsten this ratio varies from 0.464 at $T=1000°K$ to 0.440 at $T=3000°K$ (Forsythe and Worthing, Astrophys. J. **61,** 126 (1925)).

The *brightness* of the material may be compared with that of a standard lamp for a given wave-length. Usually the determination is made for red light ($\lambda = 0.665\mu$). The temperature thus determined is known as the brightness temperature, T_s, and it is *lower* than the true temperature by an amount which increases with decrease in e_λ, the spectral emissivity (for this wave-length) at any given temperature, and also increases with the temperature. Thus for tungsten, e_λ for $\lambda = 0.665\mu$ varies from 0.456 at $T = 1000°K$ to 0.415 at $T = 3000°K$ and the observed values of T_s are given in the last column of Table I.

Therefore from determinations of both T_c and T_s it is possible to deduce fairly approximate values to the true temperatures. On the other hand, it is of course possible to determine the actual value of e_λ by methods described by both Langmuir and Worthing and Forsythe, and thus to calculate T, the true temperature, from optical pyrometer measurements of T_s.

Further information on methods of determining temperatures by optical methods is given in these publications as well as in reviews by W. E. Forsythe.[15] Methods of measuring temperatures by optical methods have also been discussed by P. D. Foote and C. O. Fairchild.[16]

At temperatures below 1100–1200°K, optical methods become impracticable and it is then necessary to refer to some other property of the material such as the watts radiated per unit area or the resistance. The former method is more accurate whenever data are available on the total radiant emissivity. Naturally a correction has to be applied for the cooling effect of the leads (see below), which can be obtained by means of measurements on different lengths of the filaments. Data on the energy radiated as a function of the temperature have been summarized in the review by Lax and Pirani[17] which contains complete references to the literature. The metals mentioned in this summary are the following: W, Mo, Ta, Pt, Os, Au, Ni, Fe, C, Ag, Cu, Zr.

(1) *Tungsten*:—The temperature scale for tungsten has been determined accurately as a result of the work of I. Langmuir and H. A. Jones[13] in the Research Laboratory of the General Electric Company at Schenectady; W. E. Forsythe and A. G. Worthing[18] in the Research Laboratory of the National Lamp Works of the General Electric Company at Cleveland, and C. Zwikker[19] in the Research Laboratory of the Philips' Lamp Company in Eindhoven, Holland.

The best method for determining the temperature, as mentioned previously, is to make a vacuum lamp with a V-shaped filament from a sample of the same wire as used in the emission observations. This filament is treated during exhaust in the *same manner* as the filament in the emission tube. The

[15] W. E. Forsythe, J. Opt. Soc. Am. and Rev. Scientific Inst. **16,** 307 (1928); J. Am. Ceramic Soc. **12,** 780 (1929).

[16] P. D. Foote and C. O. Fairchild, Symposium on Pyrometry, Am. Inst. Mining and Metallurgical Engineers, **338** (1920), p. 324. See also in the same publication the papers by W. E. Forsythe (p. 291) A. G. Worthing (p. 367), and E. P. Hyde (p. 285).

[17] Lax and Pirani, Handbuch der Physik **21,** 236–240.

[18] W. E. Forsythe and A. G. Worthing, Astrophys. J. **61,** 126 (1925).

[19] C. Zwikker, Physica **5,** 249 (1925), Archives Neerlandaises des Sciences **9,** 207 (1925).

lamp is then either pyrometered against a standard lamp, or else the candle power emitted from a definite area is measured as a function of filament current and volts. Knowing the brightness (candles/cm^2) it is then possible to determine from the tables published by the above investigators the temperature as a function of filament current.

However, a very convenient method, and one which yields results almost as accurate, is to determine the watts consumed by the actual cathode used, and after correcting this for lead losses to calculate the watts/cm^2. From the tables, the corresponding temperature may then be determined quite accurately. According to Worthing and Forsythe, the energy radiated from tungsten as a function of the absolute temperature is given in watts/cm^2 by the relation.

$$\log_{10} E = 3.680(\log_{10} T - 3.3) - 1040/T + 1.900.$$

For a filament of known length, but unknown diameter, the value of $VA^{1/3}/l$ is a function of the temperature (independent of diameter), where V = voltage drop corrected for lead loss, and A is the current. Values of this function have also been tabulated by the above mentioned investigators.

(2) *Molybdenum and tantalum*: The brightness (candles/cm^2) and intensity of radiation (watts/cm^2) for seasoned molybdenum and tantalum have been determined by A. G. Worthing.[20] Either function, therefore, may be used as a measure of the temperature.

(3) *Platinum*: The spectral emissivity of well aged platinum as a function of the temperature has been determined by Worthing[21] for the wave-length $\lambda = 0.665\mu$. The value varies linearly with temperature from 0.295 at $T =$ 1200°K to 0.310 at $T = 1850$°K. The procedure used by L. A. DuBridge[22] in measuring electron emission from this metal consisted in determining the brightness temperature by comparison with a calibrated optical pyrometer in which a standard tungsten lamp was used and then correcting for the difference in emissivities of tungsten and platinum by means of Worthing's values given above.

W. Geiss[23] has determined the total radiation intensity for the metal and finds that this may be expressed as a function of the temperature by the relation

$$W = aT^n \cdot \sigma T^4 \text{ watts/cm}^2$$

where $a = 6.22 \times 10^{-4}$
$n = 0.767$
σ = Stefan-Boltzmann constant
$= 5.75 \times 10^{-12}$ watts $\cdot$ cm^{-2} deg.$^{-4}$

The total radiation emissivity from this metal as a function of temperature has also been investigated by C. Davisson and J. R. Weeks.[24]

[20] A. G. Worthing, Phys. Rev. **28,** 190 (1926).
[21] A. G. Worthing, Phys. Rev. **28,** 174 (1926).
[22] L. A. DuBridge, Phys. Rev. **31,** 236 (1928); **32,** 961 (1928).
[23] W. Geiss, Physica **5,** 203 (1925).
[24] C. Davisson and J. R. Weeks, J. Opt. Soc. and Rev. Scient. Inst. **8,** 581 (1924).

(4) *Carbon*: The spectral emissive power of graphite at $\lambda = 0.660\mu$ has been determined by C. H. Prescott and W. B. Hincke[25] for the temperature range 1250°K to 2700°K. From their data the true temperature may be calculated from values of the brightness temperature as determined by an optical pyrometer.

Assuming that carbon has an average total emissivity of 0.85, the energy radiated as a function of the temperature may also be calculated by means of the Stefan-Boltzmann law.

(5) *Nickel:* The spectral emissivity of this metal at three different wavelengths has been measured by A. G. Worthing.[26] The emissivity does not change with temperature, and for $\lambda = 0.665\mu$ the mean value 0.375 was obtained over the range 1200°K to 1650°K.

The total radiation from nickel and cobalt has been measured by C. L. Utterback.[27] The following equations were derived as expressions for the relation between energy and temperature. For nickel,

$$E = C_1 T^{5.29} \text{ for } 650° < T < 1400°\text{K}$$

$$E = C_2 T^{4.75} \text{ for } 1450° < T < 1600°\text{K}$$

For cobalt

$$E = C_3 T^{5.20} \text{ for } 672° < T < 1320°\text{K}$$

$$E = C_4 T^{4.62} \text{ for } 1380° < T < 1590°\text{K}$$

The exact values of the constants are not given.

(6) *Oxide coated filaments.* The temperature scale for these filaments differs both with the treatment and composition of the surface layer and with the nature of the coated material. Therefore this point is more conveniently treated in connection with the discussion of the emission data.

E. Lead loss correction. Owing to cooling effect by the leads, the length of filament from which the emission is obtained is less than the actually measured length. The problem of correcting for this effect of leads was first discussed by I. Langmuir,[28] and then by A. G. Worthing.[29] The paper by W. E. Forsythe and A. G. Worthing[18] gives tables from which the effect of temperature distribution along a filament may be calculated. A mechanical method for solving the differential equation involved in such problems has been developed by V-Bush and K. E. Gould.[30] However, a most comprehensive discussion of the whole problem has been published in a recent paper by I. Langmuir, S. MacLane and K. B. Blodgett,[31] and the following remarks represent a summary of the relations derived in this paper.

We consider a filament fastened to two leads which are presumably at a low temperature as compared with the maximum temperature T_m, at the

[25] C. H. Prescott and W. B. Hincke, Phys. Rev. **31,** 130 (1928).
[26] A. G. Worthing, Phys. Rev. **28,** 174 (1926).
[27] C. L. Utterback, Phys. Rev. **34,** 785 (1928).
[28] I. Langmuir, Trans. Faraday Soc. **17,** 21 (1921).
[29] A. G. Worthing, Jr. Frank. Inst. **194,** 597 (1922).
[30] V. Bush and K. E. Gould, Phys. Rev. **29,** 337 (1927).
[31] I. Langmuir, S. MacLane and K. B. Blodgett, Phys. Rev. **35,** 478 (1930).

center of the filaments. If the whole filament were at this temperature, the observed value of any property (e.g., watts radiated, electron emission), which we shall denote by H, would be H_m. The ratio $f = H_m/H$ therefore gives the correction factor by which the observed value has to be multiplied to correct for the effect of leads. Let ΔH denote the decrease in H due to one lead. Then

$$f = \frac{H_m \cdot}{H_m - 2\Delta H} \cdot \tag{4}$$

This decrease may be represented by a volt equivalent ΔV_H, and if ΔV denotes the actual decrease in voltage drop due to a single lead, it may readily be shown that

$$f = \frac{V + 2\Delta V}{V + 2\Delta V - 2\Delta V_H} \cdot \tag{5}$$

The magnitude of ΔV_H obviously depends upon the manner in which H varies with T, and also upon the ratio of the maximum temperatures T_m to that of the junction between filament and lead, T_0.

The accurate evaluation of these corrections is, therefore, a somewhat tedious matter. However, in many cases only approximate results are desired and for this purpose the following empirical type of equation may be used:

$$\Delta V_H = P(T_m/1000) - Q(T_0/1000) - R.$$

The values of the constants, P, Q and R for pure tungsten are given for different properties in Table II.

TABLE II. *Constants in formula for lead loss correction for tungsten filaments.*

H	P	Q	R	Range T_m	Range T_0
Voltage*	0.154*	0.081*	0.056*	1000–2500	any value
Candle Power	.338	.182	−.004	600–3500	300–1400
Electron Emission	.440	.158	.072	1000–3500	300– 900
Watts Radiated	.293	.160	.084	1100–3000	300– 900

* For voltage (watts input), a term, $-2.1 \times 10^{-8}\, T_0 \cdot T_m$, is to be added to the right-hand side of the above equation.

For electron emission from tungsten filaments covered with monatomic films of more electropositive elements, the values of ΔV_H will be different from those given in the table, but may be calculated by methods indicated in the original paper. For other materials than tungsten, which obey the Wiedemann-Franz law, the lead-loss correction may also be calculated approximately.

However, it would seem best in most cases either to use the guard ring principle and thus eliminate the necessity for lead-loss corrections completely, or else to measure emission on two different lengths of filaments over the same range of temperature values.[32]

[32] As, for instance, in the paper by S. Dushman and J. Ewald, Phys. Rev. **29**, 857 (1927).

F. Corrections for effect of anode voltage (Schottky effect). Equations (1) and (2) for electron emission were derived on the assumption that the field strength at the cathode is zero. Actually it is observed that the emission increases with applied anode voltage and it is therefore necessary, in order to determine the emission constants accurately, to apply some correction factor which depends upon the voltage.

An explanation of the effect of field strength on emission was first suggested by W. Schottky in 1914,[33] and hence the phenomenon is generally known as the *Schottky Effect.* As this theory has been discussed very fully by K. T. Compton and I. Langmuir,[34] it is sufficient to state that on the basis of the "electric image" theory Schottky derived a relation of the form

$$i_v = i_0 \epsilon^{(4.39/T)(dV/dx)^{1/2}} \tag{6}$$

where i_v and i_0 represent the emission at anode voltages V and 0 respectively while dV/dx denotes the voltage gradient at the emitting surface.

While this gradient can be calculated in certain simple cases from the geometrical arrangement of anode and cathode, it is much more convenient in practice to plot log i_v against $V^{1/2}$. From this plot, which should be linear if relation (6) is valid, the value of i_0 can then be determined accurately. It will be observed also that from equation (6) it follows that

$$2.303T\frac{\Delta \log_{10} i}{\Delta V^{1/2}} = 4.39S^{1/2} \tag{7}$$

where S is a constant for a given geometrical arrangement of electrodes. Hence it is possible from the plot of log i versus $V^{1/2}$ to calculate S.

Combining Schottky's equation (6) with (2) it follows that

where
$$\left.\begin{aligned} I_V &= AT^2\epsilon^{-(b_0-\Delta b_0)/T} \\ \Delta b_0 &= 4.39(dV/dx)^{1/2} = 4.39(VS)^{1/2} \end{aligned}\right\} \tag{8}$$

and I_V = electron emission per cm^2 at anode voltage V.

The following table[35] gives values of Δb_0 for $V = 100$ for the case of a filament in the axis of a cylindrical anode. The values of filament and anode diameters are such as to give field strengths, at $V = 100$, which are of the same

TABLE III. *Values of Δb_0 for $V = 100$.*

Diameter of filament	Diameter of anode	dV/dx (volts/cm)	Δb_0
0.001 cm	1.00 cm	2.96×10^4	747
.005	0.50	8.69×10^3	409
.005	1.00	7.55×10^3	381
.01	1.00	4.34×10^3	289
.1	1.00	8.69×10^2	129

[33] W. Schottky, Phys. Zeits. **15,** 872 (1914), also Zeits. f. Physik **14,** 63 (1923).
[34] K. T. Compton and I. Langmuir, Reviews of Modern Physics **2,** 124 (1930).
[35] S. Dushman and J. Ewald, G. E. Review **26,** 154 (1923).

order of magnitude as those which may be expected in ordinary thermionic devices or in experimental tubes.

While Schottky himself showed that equation (6) is valid over quite a range of temperatures, the relation was applied to derive emission data at zero field strength by S. Dushman and his associates[36] to filaments of tungsten, molybdenum and tantalum, by C. Zwikker[37] to tungsten, and by L. A. DuBridge[38] to platinum. In all these investigations a linear relation was observed between log i and $V^{1/2}$, with values of Δb_0 which could be expected from the geometrical arrangement of electrodes. W. S. Pforte[39] also confirmed the validity of Schottky's equation for a tungsten filament in the axis of a cylindrical anode (where dV/dX could be calculated directly) up to an anode potential of 2000 volts, while N. A. de Bruyne[40] has checked relation (7) quantitatively for field strengths as high as 2.5×10^6 volts/cm.

It would appear, however, as a result of the investigations of J. Becker and D. Mueller[41] that for thoriated tungsten filaments and coated filaments in general, Schottky's theory is not valid for low anode voltages, and that consequently it is not correct to use equations (6) and (7) to extrapolate to zero field strength from observations at anode potentials above 100 volts. N. B. Reynolds[42] has observed, for the case of thoriated tungsten filaments, that while log i varies linearly with $V^{1/2}$ for voltages above 100 (corresponding to field strengths greater than 10^4 volts/cm) the emission decreases much more rapidly as the field strength is decreased below this value.

A discussion of the possible explanations for these observations in the case of monatomic films as emitters has been given by Compton and Langmuir in their review.[34]

It should be observed here that for very high field strengths (in excess of 10^6 volts/cm) emission is observed from filaments even at room temperatures. This "auto-electronic" emission or cold cathode emission does not follow Schottky's relation and is virtually independent of temperature. The observations and their theoretical interpretation are discussed in a subsequent section.

In the following summary of emission data, the values of the emission constants for pure metals have been calculated, whenever possible, from data which have been corrected for the effect of field strength by means of Schottky's equation. On the other hand, in the *case of composite surfaces, the emission constants*, except where otherwise stated, *must be regarded merely as reference values* from which the emission may be calculated for anode potentials exceeding 100 volts by applying equation (6). Such emission constants cannot represent, therefore, the true values of the emission constants. It is to be hoped that in the near future further investigations will

[36] S. Dushman, H. N. Rowe, J. Ewald and C. A. Kidner, Phys. Rev., **25**, 338 (1925).
[37] C. Zwikker, Arch. Neérland, **9**, 207 (1925).
[38] L. A. DuBridge, Phys. Rev. **32**, 961 (1928).
[39] W. S. Pforte, Zeits. f. Physik **49**, 46 (1928).
[40] N. A. de Bruyne, Proc. Roy. Soc. A., **120**, 423 (1928).
[41] J. Becker and D. Mueller, Phys. Rev. **31**, 431 (1928).
[42] N. B. Reynolds, Phys. Rev. **35**, 158 (1930).

help to clear up this question[55] and thus make it possible to derive values of the emission constants which will apply to zero field strength and will correspond, therefore, to the conditions assumed in the derivation of equation (2).

II. Electron Emission Data for Elementary Substances in Solid State

Table IV gives emission constants for a number of elements. The values a and b refer to Richardson's equation (1), while A and b_0 refer to equation (2). The sixth column gives values of $\Phi_0 = 8.62 \times 10^{-5}\, b_0$, the work-function, in volts, while the next two columns give values of the emission in amps/cm^2 at the corresponding values of T. The last column gives the references in the footnotes. The letter (R) denotes that this reference is also given in

Table IV. *Electron emission constants for elementary substances.*

Element	a	$b\times10^{-4}$	A	$b_0\times10^{-4}$	Φ_0	I_T	T	Reference
C	2.37×10^{6}	4.87	5.93	4.57	3.93	2.84×10^{-3}	2000	R(8)
Ca	1.74×10^{4}	3.65	0.12	3.50	3.02			R(43)
			60.2	2.60	2.24	4×10^{-3}	1100	(44)
Cs			162	2.10	1.81	2.5×10^{-11}	500	(45)
Hf			14.5	4.10	3.53	2.8×10^{-4}	1600	(46b)
Mo		5.26	60.2	5.15	4.44	1.6×10^{-3}	2000	(47)
		5.35	60.2	5.09	4.38	2.34×10^{-3}	2000	(46a)
				3.96	3.48			(48)
	2.1×10^{7}	5.00	61	4.74	4.08	1.3×10^{-2}	2000	R(49)
	1.1×10^{8}	5.36	323	5.10	4.39	1.12×10^{-2}	2000	R(50)
Ni	4.61×10^{6}	3.40	26.8	3.21	2.77			R(51)
Pt			1.7×10^{4}	7.25	6.27	9.2×10^{-10}	1600	(52)
Ta		4.98	60.2	4.72	4.07	1.38×10^{-2}	2000	(47)
	8.32×10^{6}	5.08	29.5	4.85	4.18	3.2×10^{-3}	2000	(53)
	1.12×10^{7}	5.00	34.0	4.70	4.04	7.0×10^{-3}	2000	R(42)
Th			60.2	3.89	3.35	4.3×10^{-3}	1600	(46a)
W			60.2	5.240	4.52	1×10^{-3}	2000	(average)
		5.51	60.2	5.236	4.52	1×10^{-3}	2000	(54)
		5.53	60.2	5.250	4.53	9.1×10^{-4}	2000	(47)
		5.77	60.2	5.225	4.50	1.12×10^{-3}	2000	(46a)
	1.05×10^{7}	5.30	43.6	5.100	4.40	1.45×10^{-3}	2000	(48)
Zr			330	4.79	4.13	8.5×10^{-5}	1600	(46b)

[43] F. Horton, Phil. Trans. **A207**, 149 (1907).

[44] S. Dushman, Phys. Rev. **21**, 623 (1923).

[45] K. H. Kingdon, Phys. Rev. **25**, 892 (1925).

[46a] C. Zwikker, Proc. Amst. Acad. Sciences **29**, 792 (1926). Used his emission data to calculate A and b_0.

[46b] Phys. Zeits. **30**, 578 (1929).

[47] S. Dushman, Rowe, Ewald and Kidner, Phys. Rev. **25**, 338 (1925).

[48] M. J. Martin, Phys. Rev. **33**, 991 (1929).

[49] I. Langmuir, Trans. Am. Electrochem. Soc. **29**, 125 (1916).

[50] I. Stoekle, Phys. Rev. **8**, 534 (1916).

[51] W. Schlichter, Ann. d. Physik **47**, 573 (1915).

[52] L.A.DuBridge, Phys. Rev. **32**, 961 (1928).

[53] H. J. Spanner, Ann. d. Physik **75**, 609 (1924).

[54] C. Davisson and L. H. Germer, Phys. Rev. **20**, 300 (1922).

[55] The recent investigations of W. B. Nottingham (Phys. Rev. **35**, 1126 (1930) promise to contribute considerable information on this topic.

Richardson's book, "Emission of Electricity from Hot Bodies." The data are the same, with slight changes or additions, as those originally published by the writer in International Critical Tables, 6, 53–57 (1930), and the following remarks on the relations between the emission constants and on conversion factors are taken from the introductory statement to the tables giving these emission data.

As pointed out by Richardson, it is impossible, within the range of temperatures (T_1 to T_2) over which emission data can be observed, to distinguish between equations (1) and (2). In this range

$$\frac{b_1 - b_0}{2.303} = \frac{3 \log_{10}(T_1/T_2)}{2(1/T_2 - 1/T_1)} \tag{9a}$$

or approximately

$$b_0 = b - 1.5\frac{(T_1 + T_2)}{2} = b - 1.5T_A \tag{9b}$$

and

$$\log_{10}(a/A) = 1.5 \log_{10} T_A + (b - b_0)/(2.303T_A). \tag{10}$$

In these equations T_1 and T_2 represent the upper and lower values of the range of temperatures over which the emission data are recorded, and T_A therefore represents the average value. Wherever only the values of a and b (or Φ) are available, equations (9b) and (10) have been used to calculate values of A and b_0 (or Φ_0), and both sets of values are given in the table.

The unit of $I_T = 1$ ampere/cm^2 $= 3 \times 10^9$ esu/(sec. cm^2) $= 0.1$ emu/(sec. cm^2) $= 6.28 \times 10^{18}$ electron/(sec. cm^2), while a and A are given in amps/cm^2 deg$^{1/2}$ and amps/cm^2 deg^2 respectively.

Notes on emission constants for solids. In general most data obtained before 1913–1914 are quite unreliable and have been omitted to a large extent from the table. References to these older data may be obtained in Richardson's book.

Tungsten. The most satisfactory data are those obtained by C. Davisson and L. H. Germer,[54] C. Zwikker[46] and S. Dushman and his associates.[47] The agreement between the three sets of results is as good as can be expected, and the average values given in the table may be regarded as probably the most accurate emission data obtained for any substance.

Molybdenum. The values of A and b published by the writer in 1925[47] have been corrected for the revised temperature scale published more recently by A. G. Worthing. They are in good accord with Zwikker's data,[46a] and while the latter gives as most probable value, $A = 65$, the results are in satisfactory agreement with the value $A = 60.2$.

The value $b = 39{,}600$ was obtained by M. J. Martin[48] on a thoroughly degassed sample of molybdenum from emission data for the range 1300°K to 1600°K. While this value is in good agreement with the work function determined for the same metal from the long wave limit for the photoelectric effect, it is abnormally low compared with any of the other published values.

A possible explanation would be the presence of traces of alkaline earth metals which would not be eliminated by treatment at 1600°K.

For comparison there are also given the less accurate values obtained by Langmuir[49] and Stoekle.[50]

Tantalum. The most reliable data are those observed by S. Dushman and his associates.[47] It will be observed that these results do not differ greatly from those obtained by Langmuir[8] in 1913, and the difference may be readily accounted for by differences in temperature scale. Spanner's[53] original emission data are not given in his paper, but merely plots of log (I/T^2) versus $1/T$. These were enlarged and used to derive the values of A and b_0 given in the table. For this reason, and also because, apparently, no correction was made for end losses or Schottky effect, these values are probably only approximate.

Platinum. A great many investigations have been carried out on the emission constants from this metal. The values of the constants obtained by these different investigators as tabulated by L. A. DuBridge[56] together with some additional values are given in Table V. The values given in Table IV obtained by DuBridge as the result of a prolonged de-gassing of the metal under extremely good vacuum conditions undoubtedly represent the most accurate data obtained for this metal.

TABLE V. *Concurrent values of emission constants for platinum.*

Author	Reference	A	$b_0 \times 10^{-4}$	Φ_0
Wilson	R(57)	10.7×10^{-4}	2.56	2.18
Lockrow	(58)	1.36×10^{-2}	4.40	3.87
Richardson	R	53.8	4.73	4.08
Richardson	R	39.8	4.75	4.10
Suhrmann	(59)	30.1	4.85	4.18
Schlichter	R(51)	51.6	4.91	4.23
Lockrow	(58)	4.46×10^{-2}	5.00	4.31
Du Bridge	(56)	11.5	5.45	4.69
Deininger	R(60)	16.2	5.85	5.02
Horton	R(43)	11.5	5.90	5.08
Spanner	(53)	28.8	6.06	5.22
Wilson	R(57)	366	6.30	5.45
Du Bridge	(56)	235	6.50	5.60
Richardson	R	265	6.55	5.65
Wilson	R(57)	620	7.00	6.00
Du Bridge	(56)	6450	7.20	6.20
"		8130	7.30	6.30
"		14000	7.42	6.40
Langmuir	(49)	1.45×10^{7}	7.80	6.71

It should be pointed out, however, that in spite of the thorough degassing of the metal, it may still be questioned whether the surface was thoroughly free of adsorbed oxygen. As shown by Langmuir,[61] oxygen condenses readily

[56] L. A. DuBridge, Phys. Rev. **31**, 736 (1928).
[57] H. A. Wilson, Phil. Trans. **208A**, 251 (1908).
[58] L. Lockrow, Phys. Rev. **19**, 97 (1922).
[59] R. Suhrmann, Zeits. f. Physik **13**, 17 (1922).
[60] F. Deininger, Ann. d. Physik **25**, 285 (1908).
[61] I. Langmuir, J. Am. Chem. Soc. **38**, 2221 (1916), and other papers.

on *clean* platinum as a monomolecular film which is stable even at higher temperatures. Such a film would have the same effect as in the case of tungsten and oxygen, that is, a considerable decrease in emission, as shown by the increase in both A and b_0. This point is discussed further in the section dealing with effect of gases on emission.

Attention has been drawn by DuBridge[62] to the interesting fact that large values of A occur concurrently with large values of b. This is very evident from the data in Table V. The significance of this observation will be discussed in connection with that of the "constant" A in a subsequent section.

Other elements. The data for carbon are probably as reliable as can be expected in view of the lack of knowledge of an accurate temperature scale. On the other hand, the values given for calcium are probably only roughly approximate, while those for nickel correspond to emission currents which are certainly considerably greater than any actually observed by the writer on vacuum treated wires. In fact, it is quite likely that the emission from nickel, iron, and other metals of Group VIII, like that observed in the case of platinum, is extremely low, corresponding to work-functions of at least 5 volts.

Thorium, zirconium and hafnium. Earlier emission data for solid filaments of these three metals were published by C. Zwikker[46a] in 1926. In the case of thorium the observations were in quite satisfactory agreement with the value $A = 60.2$; but in the case of the other two metals very large values of A were obtained. It seemed possible to account for these results by the presence of oxide films which could not be volatilized readily from the emit-

TABLE VI. *Concurrent emission constants for zirconium and hafnium (Zwikker).*

Metal	A	$b_0 \times 10^{-4}$	Φ_0
Zirconium	37,200	6.06	5.22
	9,500	5.80	5.00
	3,000	5.23	4.51
	330	4.79	4.13
Hafnium	55,000	5.95	5.13
	6,600	5.69	4.90
	1,000	5.03	4.34
	70	4.39	3.78
	145	4.10	3.53

ting surfaces. This suggestion is in agreement with observations made K. H. Kingdon, S. Dushman and others in the case of oxide films on tungsten, molybdenum, and tantalum that such films yield large values of A and also abnormally high values of b_0. More recently Zwikker[46b] has published further data on the emission from zirconium and hafnium which apparently are in agreement with these views. Table VI gives Zwikker's values for the emission constants as observed during successive stages of ageing of filaments. The final values are those given in Table IV as probably the most accurate.

[62] L. A. DuBridge, Proc. Nat. Acad. **14**, 788 (1928).

III. Electron Emission from Monatomic Films

A. Thorium on tungsten.

(1) *Thermal treatment of thoriated tungsten filaments.* During the course of development of ductile tungsten wire for use in the lamp industry it was discovered that a small addition of thorium oxide (about 0.5 to 0.75 percent by weight) prevents offsetting of the filaments when operated on alternating currents. In 1913 I. Langmuir and W. Rogers[63] discovered that by proper thermal treatment of such wire a very much higher electron emissivity could be obtained than that observed from filaments of pure tungsten. Subsequently Langmuir investigated more thoroughly the conditions under which this high electron emission may be obtained and developed a comprehensive theory of the phenomena involved in the activation and deactivation of thoria-doped tungsten filaments.[64] At the present time the amount of thoria added to tungsten for use as a hot cathode varies between one and two percent. It should be observed here that the magnitude of the saturation emission obtained by the thermal treatment described below is independent of the actual thoria content, but for practical purposes, such as increased rate of activation and longer life under operating conditions, it has been found advantageous to have the thoria content as large as practicable.

The procedure adopted in obtaining thorium emission consists of two operations:

(a) By flashing the filament for a short interval (usually one or two minutes) at temperatures above 2700°K, some of the thoria is reduced to thorium. The thorium atoms immediately begin to diffuse towards the surface, but evaporate rapidly at the surface at these high temperatures.

(b) If the temperature is now lowered to about 2100–2200°K, the diffusion is still quite rapid, but the evaporation is decreased and the thorium accumulates on the surface as an adsorbed layer one atom thick.

For any temperature there is thus obtained for the density of thorium atoms on the surface an equilibrium state which depends upon the relative importance of the three factors: reduction of ThO_2; diffusion and evaporation of Th atoms. The observed emission thus depends upon the fraction θ, of the surface which is covered with thorium atoms. As has been shown by Langmuir, the rate of activation at any temperature is given by the relation

$$N_0 \frac{d\theta}{dt} = DG(1 - \theta) - E \tag{11}$$

where N_0 = number of atoms of thorium per unit area for the surface completely covered, D = diffusion constant, G = concentration gradient, and E = rate of evaporation. If the thorium atoms could accumulate on the surface as a layer more than one atom thick, the rate of diffusion would be given by DG. But the actually observed form of relation for the rate of acti-

[63] I. Langmuir and W. Rogers, Phys. Rev. **4**, 544 (1914).
[64] I. Langmuir, Phys. Rev. **22**, 357 (1923).

vation shows that the thorium atoms evaporate much more rapidly from underlying thorium atoms than when in direct contact with the tungsten atoms. Hence there is an "induced evaporation" whose rate is given by $DG\theta$. This type of evaporation is to be carefully distinguished from the normal temperature evaporation of thorium atoms when in contact with tungsten atoms, the rate of which is denoted by E, and which is much lower at any given temperature than that of the induced evaporation.

The ranges of temperatures at which the different types of reaction predominate in thoriated tungsten filaments are found to be as follows:

(i) $T \geqq 2600°K$. *Reduction range.*

Marked reduction of ThO_2 occurs, accompanied by extremely rapid rates of diffusion and evaporation. The equilibrium value of θ is approximately 0, i.e., the surface is in the completely inactive condition.

(ii) $T = 2300–2600°K$. *Deactivating range.*

The lower the temperature, the lower the rate of formation of fresh Th by reduction, and this becomes negligibly small at the lower limit of temperatures. In this range of temperatures, the rate of normal evaporation is greater than the rate of diffusion, so that an activated surface becomes deactivated, and tends to an extremely low equilibrium value.

(iii) $T = 2000–2300°K$. *Activating range.*

In this range of temperatures the rate of diffusion usually exceeds that of evaporation and θ tends to approach the value 1 as the temperature is decreased to the lower limit.

(iv) $T = 1800°–2000°K$. *Operating range.*

Rate of evaporation and that of diffusion are so low that the value of θ changes only very slowly or not at all over prolonged periods of time. Ordinarily thoriated tungsten filaments are operated commercially at about 1900°K.

(2) *Value of* N_0. From considerations based on the size of the unit crystal lattice for both tungsten and thorium, Langmuir concludes that

$$N_0 = 7.56 \times 10^{14} \text{ atoms per cm}^2$$

(3) *Rate of production of metallic thorium.* The heat of the reaction involved in the reduction of ThO_2 by W is $Q_r = -138{,}000$ cal. per gram-atom. For a filament of 0.00389 cm radius and containing 1 percent oxide (corresponding to 4.4×10^{20} ThO_2 molecules per cm^3) the average concentrations of metallic thorium in parts per million obtained at different temperatures are as follows:

1800°	1900°	2000°	2100°	2200°	2400°	2800°	3000°K
1.4	2.8	5.0	8.4	14	32	118	200

(4) *Rate of diffusion of thorium atoms.* The diffusion constant D (in cm^2 $sec.^{-1}$) as a function of temperature is given, according to Langmuir's observations, by

$$\log_{10} D = 0.044 - 20540/T \qquad (12)$$

corresponding to a heat of diffusion, $Q_0 = 94{,}000$ cal. per gram-atom.

Subsequent investigations in the Research Laboratory at Schenectady have shown, however, that the rate of diffusion of thorium in thoriated tungsten filaments is governed to a great extent by the crystal structure of the metal. Consequently the value given above must be regarded as valid only for the type of thoriated wire used at that time (1915–1922). For thoriated wire as used at present the values of D, as calculated by the methods developed by Langmuir, apparently vary with the diameter of the wire and also with the mode of temperature treatment during manufacture. The more detailed discussion of these phenomena and their interpretation will be given in a separate paper by N. B. Reynolds and the writer.

(5) *Normal evaporation of thorium.* The rate of evaporation, E, in atoms per cm^2 per sec. is given by the relation

$$\log_{10} E = 31.434 - 44500/T. \tag{13}$$

The heat of evaporation of Th atoms from tungsten is therefore $Q_E = 204{,}000$ cal. per gram-atom. This is higher than the value of the heat of evaporation for tungsten from tungsten, which, according to Langmuir and Jones,[13] is 191,880 cal. per gram-atom.

This value of E applies only for the range $\theta = 0.1$ to $\theta = 0.85$ approximately. For values of $\theta < 0.1$, E decreases, while for values of $\theta > 0.85$, it increases.

(6) *Variation in emission constants with* θ. The fraction θ is defined according to Langmuir by the relation

$$\theta = \frac{b_\theta - b_w}{b_{Th} - b_w} \tag{14a}$$

where b_θ, b_{Th} and b_W represent the values of b_0 in the emission equation for the partly covered surface, completely covered surface and pure tungsten respectively.

This may evidently be expressed in the form

$$b_\theta = b_w + C\theta \tag{14b}$$

where $C = (b_{Th} - b_W)$.

If $\log A_\theta$ is a linear function of b_θ, it also follows that

$$\theta = \frac{\log I_\theta - \log I_w}{\log I_{Th} - \log I_w} \tag{15}$$

where I_θ, I_{Th} and I_W are the values of the electron emission per unit area at the same temperature for the three surfaces. This relation is obviously more convenient to use in practice, and although the values of $\log A_\theta$ for thoriated tungsten, as will be shown below, do not vary linearly with θ over the whole range, equation (15) leads to values of θ which are not very different from those obtained by the use of the more rigorous equation (14a).

Table VII gives values of A_θ and b_θ for different states of activation as observed by Dushman and Ewald.[65] The values of θ were calculated by

[65] Dushman and Ewald, Phys. Rev. 29, 857 (1927).

means of equation (14a) using the values $b_{Th}=30{,}500$ and $b_W=52{,}400$. The last two columns give values of I_θ for $T=1900°K$, and of Φ_0.

In this table (as in nearly every other table of emission data) the values of the emission constants have been derived from measurements carried out at anode voltages of 100 to 500 by applying Schottky's equation (6). Consequently, in view of the remarks made in a previous section, these constants do not represent the correct values for zero field strength.

The values of A_θ and b_θ for $\theta=0.18$ are to be regarded as doubtful. On the whole, the observed values point to the conclusion that from $\theta=0.95$ to $\theta=0$, log A_θ varies linearly with b_θ and therefore with θ, while a minimum value occurs for A_θ at $\theta=0.05$.

TABLE VII.

θ	A_θ	b_θ	I_θ	Φ_0
1	3.0	30,500	1.166	2.63
0.95	1.50	31,460	0.349	2.71
.83	2.08	34,150	.118	2.94
.72	3.74	36,570	.0594	3.15
.56	7.76	40,070	.0195	3.45
.43	10.86	42,840	.0064	3.69
.25	15.81	47,050	.0010	4.06
.18	(8.2)	48,360		4.17

K. H. Kingdon[66] has also published values of emission constants for thoriated tungsten filaments. For the relation between A and θ he has suggested an empirical equation of the form

$$A_\theta = a_1{}^\theta + a_2{}^{1-\theta} - 1 \text{ amps/(cm}^2 \text{ deg.}^2) \qquad (16)$$

where a_1 gives the value of A for $\theta=1$; and a_2 for $\theta=0$. For the case of thorium on tungsten he finds $A=7$ and $b_0=31{,}200$, while for tungsten he uses the values $A=60$ and $b_0=52{,}200$. As pointed out by Kingdon, "The artificial nature of expression (16) makes it seem likely that this is not the true expression for A_θ, but merely one which approximates to the true expression in the cases studied."

TABLE VIII. *Values of I as a function of θ for thoriated tungsten for T=1900°K.* $\log_{10} I_\theta = 0.0667 - 3.7088(1-\theta)$

θ	I (amps/cm^2)	θ	I	θ	I
1.0	1.166	0.7	8.99×10^{-2}	0.4	5.51×10^{-3}
0.9	4.964×10^{-1}	0.6	3.83×10^{-2}	0.2	1.00×10^{-3}
0.8	2.113×10^{-1}	0.5	1.63×10^{-2}	0	2.28×10^{-4}

Richardson has, however, shown that for various adsorbed surfaces there exists a linear relation between log A and b, and, as a matter of fact, if in equation (16), a_1 is very much greater than a_2, then for all values of θ except those near $\theta=0$, log A_θ becomes a linear function of θ that is, of b_θ.

It is of interest to point out in this connection that since θ is approxi-

[66] K. H. Kingdon, Phys. Rev. **24**, 510 (1924).

mately a linear function of log I_θ, the value of the emission decreases exponentially with θ. Table VIII shows the variation in I_θ with θ, calculated by means of equation (15) for $T = 1900°K$, which is the approximate temperature for commercially operated thoriated tungsten cathodes

The effect of accelerating fields on the work-function from thoriated tungsten has been investigated by W. B. Nottingham.[67] From his measurements he has deduced the conclusion that for *zero field*

$$A = 59, \quad b_0 = 36,500 \quad \Phi_0 = 3.15 \text{ volts}$$

when maximum emission was obtained, and

$$A = 56, \quad b_0 = 36,000 \quad \Phi_0 = 3.10 \text{ volts}$$

for a thorium layer *thicker* than that required to give maximum emission. This result, he points out, is in agreement with Langmuir and Kingdon's observation that the contact potential between Th—W and W is 1.46 volts, since subtracting this value from the work function for W (4.53 volts) gives 3.07 volts.

TABLE IX. *Emission, emission efficiency and other constants for thoriated tungsten* ($\theta = 1$)

T	$I\left(\frac{\text{amps}}{\text{cm}^2}\right)$	$\frac{I}{W}\left(\frac{\text{amps}}{\text{watt}}\right)$	D (cm² sec.$^{-1}$)	$E\left(\frac{\text{atoms}}{\text{cm}^2 \text{ sec.}}\right)$	I
1000	1.73×10^{-7}	2.87×10^{-7}			8.01×10^{-9}
1100	3.31×10^{-6}	3.22×10^{-6}			2.69×10^{-7}
1200	3.95×10^{-5}	2.38×10^{-5}			5.33×10^{-6}
1300	3.27×10^{-4}	1.27×10^{-4}			6.21×10^{-5}
1400	2.03×10^{-3}	5.30×10^{-4}	2.36×10^{-15}	0.445	5.36×10^{-4}
1500	1.00×10^{-2}	1.81×10^{-3}	2.24×10^{-14}	58.5	3.51×10^{-3}
1600	4.06×10^{-2}	5.24×10^{-3}	1.61×10^{-13}	4.18×10^{3}	1.83×10^{-2}
1700	1.40×10^{-1}	1.32×10^{-2}	9.15×10^{-13}	1.81×10^{5}	7.91×10^{-2}
1800	4.28×10^{-1}	3.09×10^{-2}	4.29×10^{-12}	5.15×10^{6}	2.93×10^{-1}
1900	1.164	6.24×10^{-2}	1.71×10^{-11}	1.03×10^{8}	9.50×10^{-1}
2000	2.864	1.19×10^{-1}	5.94×10^{-11}	1.53×10^{9}	2.75
2100			1.83×10^{-10}	1.75×10^{10}	
2200			5.10×10^{-10}	1.61×10^{11}	
2400			3.06×10^{-9}	7.80×10^{12}	
2600			1.39×10^{-8}	2.08×10^{14}	
2800			5.11×10^{-8}	3.48×10^{15}	

Table IX gives values of the emission and emission efficiency for completely activated surfaces and also values of D and E according to Langmuir, as functions of the temperature. Since at temperatures exceeding 2000° K evaporation becomes so rapid that the maximum value of θ can no longer be obtained, values of I and I/W for such temperatures are not given in the table.

These values have been calculated from the values of $A = 3.0$ and $b_0 = 31,500$. The last column gives values of I calculated for zero field by means of Nottingham's constants, $A = 59$, $b_0 = 36,500$.

Finally reference should be made to Table VIII in Langmuir's original paper for data on the values of G_r, the concentration gradient, θ_r, the equilibrium value of θ, the electron emission, and life and time of activation as

[67] W. B. Nottingham, Phys. Rev. **36**, 386 (1930).

functions of T for a filament 0.00389 cm in diameter and containing 1 percent ThO_2.

Attention may also be drawn to the differences between the values for A and b_0 for thoriated tungsten and those given in Table IV for metallic thorium, as observed by Zwikker. Thus it is observed that *the electron emission for solid thorium is less than that for a monatomic film on tungsten.*

B. Monatomic films of other rare earth and alkaline earth metals adsorbed on tungsten and molybdenum.

If other rare earth oxides are used instead of ThO_2 in "doping" the tungsten, and the filaments thus obtained are treated thermally, emission phenomena are obtained which resemble those observed for thoriated tungsten. The behavior of such filaments has been investigated by S. Dushman, D. Dennison and N. B. Reynolds,[68] and their observations may be summarized briefly as follows:

The monatomic films produced by reduction of the oxide at high temperature and subsequent activation at lower temperatures are not as stable as those obtained with thoria, owing both to more rapid diffusion and evaporation. This makes it difficult to obtain accurate emission data for completely covered surfaces. The values of A and b_0 obtained for different films with filaments 0.01 cm diameter are given in Table X along with values of D, the diffusion constant in cm^2 $sec.^{-1}$ at $T=2000$, Q_D, the heat of diffusion (gram-calories per gram-atom), and E, the rate of evaporation in atoms/cm^2 sec. at 2000°K.

TABLE X. *Emission constants, rates of diffusion and evaporation for films of rare earth metals on W and Mo.*

Emitter	A	b_0	$D\times10^{11}$	Q_D	$E\times10^{9}$
Ce-W	8.0	31,500	95	83,000	1450
La-W	8.0	31,500	—	—	—
U-W	3.2	33,000	1.3	100,000	>Th
Yt-W	7.0	31,300	324	78,000	68
Zr-W	5.0	36,500	1820	62,000	—
Th-W	3.0	30,500	5.9	94,000	1.53
Th-Mo	1.5	30,000	102,000	52,000	5400

The emission phenomena for films of *thorium on molybdenum* have been investigated by N. B. Reynolds and S. Dushman.* In the table are given values of the emission constants and also values of D and E at $T=2000$. While the values of A and b_0 are lower than those obtained for films of tungsten, the actual emission per watt is approximately the same owing to the lower radiant emissivity of molybdenum. On the other hand, owing to the extremely large values of both D and E, θ decreases rapidly below unity for values of $T>1700$°K. For the heat of evaporation of thorium from molybdenum the value obtained was $Q_E=170{,}000$ cal./gm-atom (approximately).

[68] S. Dushman, D. Dennison and N. B. Reynolds, Phys. Rev. **29**, 903 (1927).

* Unpublished data.

Comparing the emission constants given above for monatomic films of zirconium with those given in Table IV for pure zirconium filaments, it is observed that as in the case of thorium, *the emission from monatomic films is considerably greater than that observed from the solid metal.*

In connection with investigations on the emission from oxide-coated cathodes there have also been made certain observations on the emission from monatomic films of barium on tungsten and oxidized tungsten. While the phenomena observed with these films are discussed in detail in a subse-

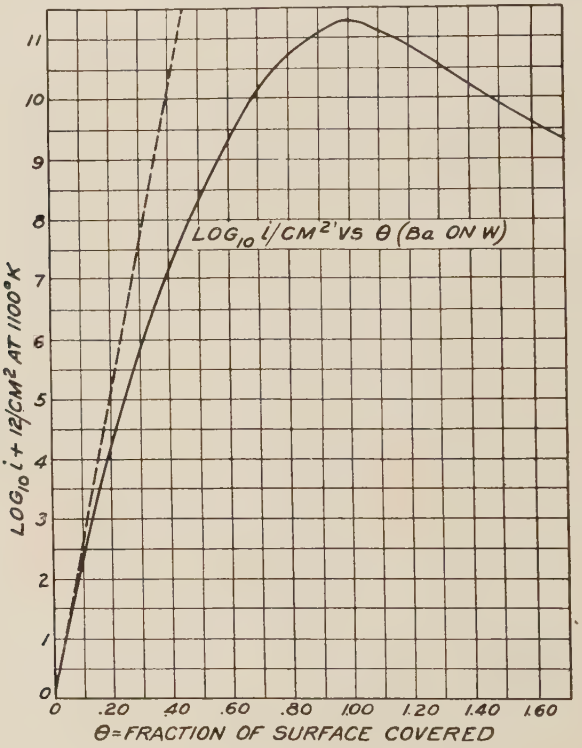


Fig. 1. Variation of electron emission with fraction of tungsten surface covered with barium (J. A. Becker[70]).

quent section, it is of interest to mention briefly the results obtained by J. A. Becker.[69] The variation in emission with θ for a film of barium on tungsten as observed by him is shown in Fig. 1.[70] For such films he finds that under the conditions $0 < \theta < 0.85$,

$$\log (i/i_0) = 1.1(\log i_1/i_0)(1 - \epsilon^{2\theta})$$

where i_1 and i_0 denote the values of emission for $\theta = 1$ and $\theta = 0$ respectively. Furthermore, the *decrease* in work function, $\Delta\Phi$, is given by

[69] J. A. Becker, Phys. Rev. **33,** 1082 (1929).
[70] J. A. Becker, Phys. Rev. **34,** 1323 (1929).

$$\Delta\Phi = 1200\pi N(4.77 \times 10^{-10} l) \text{ volts}$$

where N = number of barium ions per cm², and l = diameter of ion.

C. Caesium on tungsten and oxidized tungsten.

(1) *General remarks.* The emission phenomena for monatomic films of caesium adsorbed on tungsten and oxidized tungsten were first investigated by I. Langmuir and K. H. Kingdon [71,72,73]. The phenomena observed are different in many respects from those observed in the case of thorium on tungsten, mainly because the caesium is adsorbed from the vapor and hence the equilibrium value of θ depends not only on the rate of evaporation of caesium atoms from the surface (that is, on the temperature of the latter), but also on the rate at which caesium atoms strike the surface, and therefore on the pressure of the vapor, that is, the bulb temperature.

In a bulb containing caesium at 30°C the adsorbed layer on pure tungsten gives a maximum emission of about 10^{-4} amp./cm² at a filament temperature of 700°K. At higher filament temperatures the rate of evaporation increases and θ decreases with consequent decrease in emission. On the other hand, if the filament is first heated in oxygen, the presence of O atoms on the W atoms retards the evaporation of adsorbed Cs atoms and a much greater value of the maximum emission is obtained. The process of oxidation has been described by Kingdon [71] as follows:

"After flashing the filament to 2700°K in vacuum for a short time, it is heated at about 1900°K in oxygen at a pressure of about 0.02 mm for a few seconds. (Somewhat similar results are obtained if the oxygen is merely allowed to come in contact with the clean, cold filament). The excess oxygen is then pumped out, and caesium distilled in. Some caesium is presumably adsorbed on the filament, but the electron emission which may be obtained at temperatures of from 700°K to 1000°K is very small. If, however, the filament is heated for a few seconds at from 1200° to 1600°K, on returning to the lower temperatures it will be found that the emission has increased many fold and is of the order 0.35 amp./cm² at 1000°K, with the bulb at 30°C. The heating at from 1200° to 1600°K probably frees the surface of excess oxygen and of adsorbed deleterious gases such as CO. The remaining oxygen is probably held on as negative oxygen ions, and these ions adsorb the caesium more strongly than a tungsten surface does. With the bulb at 30°C the emission is at a maximum at a filament temperature of about 1000°K. At higher filament temperatures the caesium evaporates from the filament faster than it can be renewed by condensation from the vapor in the bulb, and the emission decreases."

(2) *Electron emission as a function of the pressure of caesium vapor and temperature of filaments.* Fig. 2, taken from the paper by Langmuir and Kingdon,[73] shows $\log_{10} I$ as a function of the reciprocal of the absolute tem-

[71] I. Langmuir and K. H. Kingdon, Phys. Rev. **24**, 510 (1924).
[72] Phys. Rev. **25**, 892 (1925).
[73] Proc. Roy. Soc. **A107**, 61 (1925).

perature of the tungsten filament (plotted as abscissa) for various bulb temperatures, the values of which are indicated on each curve. It is observed that each curve consists of three parts:

"(I) A low temperature region in which the emission apparently follows Richardson's equation, giving a straight line descending to the right in Fig. 2.

"(II) An intermediate region where the plot in Fig. 2 is curved, the current reaching a maximum as the temperature is raised and then falling. This region may be regarded as a transition curve between the straight portions I and III.

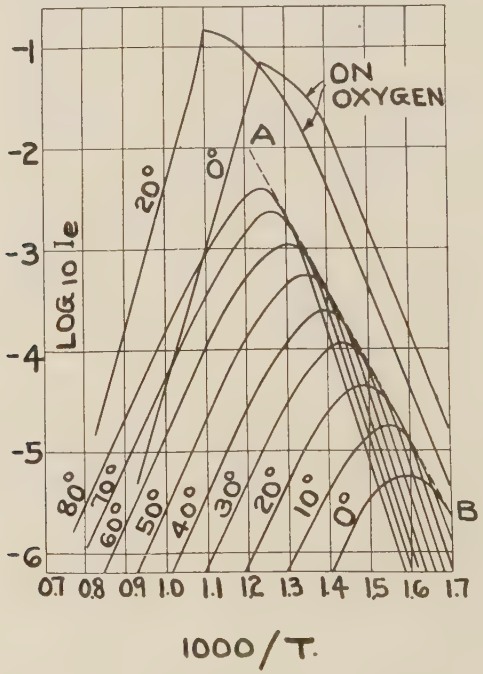


Fig. 2. Thermionic emission from tungsten in caesium vapor as a function of filament temperature for various bulb temperatures. (I. Langmuir and K. H. Kingdon[73]).

"(III) A high temperature region, where the plot in Fig. 2 becomes straight again, but the current decreases with rising temperature. An equation of the Richardson type may be used to represent these lines, but the quantity b would then have a negative value.

"Fig. 2 shows that both in Regions I and II the straight lines at the various pressures are approximately parallel to one another. The envelope of the family of curves is a straight line, shown as the dashed line AB. All the curves of the family are, in fact, very nearly alike, and differ principally in being displaced in a direction parallel to AB. These displacements are found to be proportional to the logarithms of the pressures of caesium.

"These conclusions may be summarized by saying that these electron emission data are given by the equation

$$\log I_e + M \log p = f(1/T + N \log p) \tag{17}$$

where M and N are constants and f denotes a function characteristic of the shape of the curves in Fig. 2. From the data we find that the actual values of M and N are

$$M = -0.83; N = 4.4 \times 10^{-5} \text{ deg.}^{-1}$$

"The pressures of caesium vapor used in these calculations were found by measurements of the positive ion currents obtained with filament temperatures of about 1500°K, where all the atoms that strike the filament are converted to ions. The vapor pressures obtained in this way are accurately given by the equation (p in dynes/cm^2)

$$\log_{10} p = 10.65 - 3992/T. \tag{18}$$

"From these considerations it may readily be seen that in Regions I and III the electron emission of a tungsten filament of temperature T in caesium vapor at a pressure p is expressible approximately by the equation

$$I_e = \alpha p^{\gamma} \epsilon^{-\beta/T} \tag{19}$$

where α, β and γ are constants.

"The data illustrated in Fig. 2 give the following values of the constants, when I_e is in amperes per cm^2 and p in baryes, (Table XI)."

TABLE XI. *Emission constants for Cs-W and Cs-O-W as function of vapor pressure of Cs.*

Conditions	α	β	γ
Caesium on Tungsten			
Region I	6.65×10^{14}	+31300	−0.70
Region III	8.6×10^{-12}	−19100	+1.66
Caesium on Adsorbed Oxygen			
Region I	1.04×10^{10}	+23400	−0.62
Region III	2.95×10^{-11}	−32200	+1.96

Similar observations to those shown in Fig. 2 have been made by T. J. Killian[74] for the emission from tungsten and oxidized tungsten filaments in rubidium and potassium vapors, the results being adequately represented by equation (19) with different values of the constants.

(3) *Electron emission constants for Cs-O-W filaments.* Fig. 3[71] shows the variation in emission from a Cs-O-W filament, with both bulb temperature and filament temperature, in region I where the surface is largely covered with caesium. It will be observed that a change in bulb temperature from −185° to + 60° C decreases the emission at 625°K ($10^3/T = 1.60$) eighty fold. As Kingdon points out: "The only curves from which any reliable estimates of the electron emission constants for a caesium-oxygen-tungsten filament can be made are those taken with the tube immersed in liquid air. Here there is no continual condensation of caesium on the filament, and the rate of evaporation of the absorbed layer is so small at 600 to 700°K that a temperature characteristic may be taken without changing the condition of

[74] T. J. Killian, Phys. Rev. **27**, 578 (1926).

the surface appreciably. Such runs give the values of the constants in Eq. (2) as $A = 0.003$ *amp./cm*^{2} *deg.*^{2}; $b_0 = 8300°$. As a little oxygen or tungsten was undoubtedly exposed, b_0 is a little higher and A is almost certainly much larger than the value which would characterize a complete film of caesium on oxygen on tungsten. However, as an approximation, we may take the values *for a complete caesium film to be* $A = 0.001$ *amp./cm*^{2} *deg.*^{2}, $b_0 = 8300°$" (*corresponding to* $\Phi_0 = 0.695$ *volts.*)

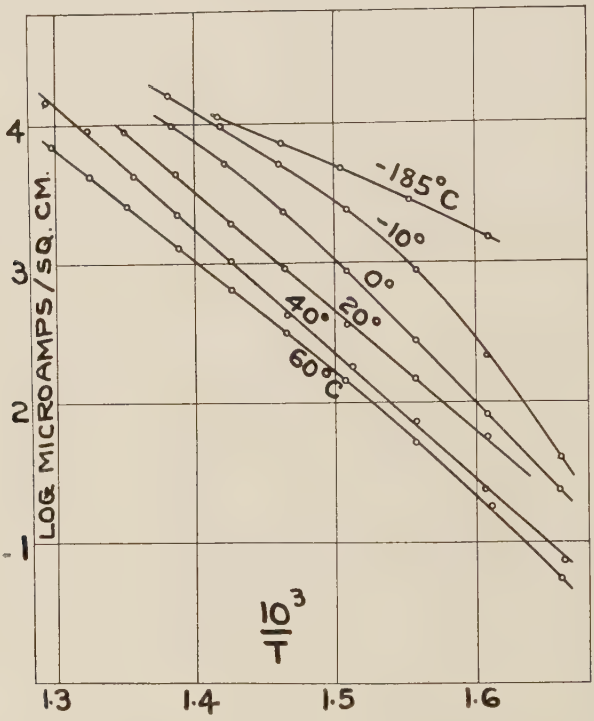


Fig. 3. Variation in electron emission from a Cs-O-W filament with both bulb temperature and filament temperature. (K. H. Kingdon[71]).

(4) *Relation between* A_θ *and* b_θ *for Cs-O-W and Cs-W filaments.* In order to account for the course of the curves shown in Fig. 3, Kingdon has used an equation of the form

$$I_\theta = [a_1^\theta + a_2^{1-\theta} - 1]A_0T^2\epsilon^{[-b_1\theta+b_2(1-\theta)]/T} \tag{20}$$

which is the result of combining his empirical relation for A_θ (equation (16) above) with equation (14) for b_θ.

For a complete film of oxygen on tungsten he has derived the values $A = 5\times10^{11}$ *amp/cm*^{2} *deg.*^{2}, $b_0 = 107{,}000°$ ($\Phi_0 = 9.22$ volts). These are to be compared with the values given above of A and b_0 for complete caesium films. The values of A and b_0 for O-W filaments incompletely covered with Cs will vary between these extremes according to equation (20), and thus lead to an explanation of the curves in Fig. 3. "At the high bulb temperatures and low

filament temperatures the surface is practically completely covered with caesium, and the emission is small on account of the small A term. As the bulb temperature is lowered, a very small fraction of the filament surface is bared of caesium and the increase in the A term increases the emission in spite of the decrease due to the increase in b. The same argument shows that at a constant bulb temperature, the slight decrease in the amount of caesium on the surface at the higher filament temperatures gives the curves too steep a slope, and this slope is not even an approximate measure of b_0 on account of the large variation in A."

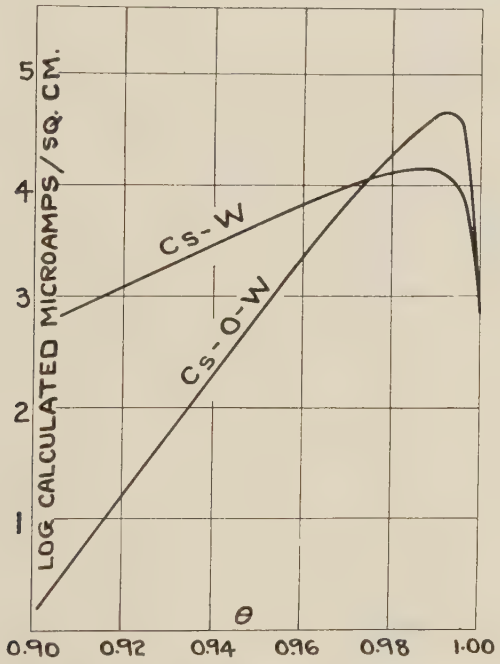


Fig. 4. Calculated variation of electron emission from Cs-O-W surface as function of fraction of surface covered. (K. H. Kingdon[71]).

With the values given above of A and b for Cs-O-W and O-W, Kingdon has calculated the electron emission for Cs-O-W filaments at 625°K as a function of θ. These values which are to be regarded as only approximate are shown in the curve in Fig. 4. Similarly for Cs-W the emission has been calculated from the values

$$A = 0.001,\ b = 8300°;\ A_w = 60.2,\ b_w = 52{,}200°.$$

It will be observed from the curves that for Cs-O-W the emission is a maximum at $\theta = 0.99$, while for Cs-W the maximum emission is obtained at about $\theta = 0.985$.

As shown by Kingdon, $\log A_\theta$ for Cs-O-W is approximately a linear function of b_θ (and therefore of θ) for values of $b_\theta > 10{,}000$, while for lower values of b_θ, $\log A$ decreases much more rapidly than b_θ.

Mention should also be made here of the emission constants obtained by L. R. Koller[75] for films of Cs on oxygen on silver (such as are used in caesium photoelectric cells). Using cathodes of 10 cm^2 area he was able to measure thermionic emission at temperatures below 473°K. The constants determined from these measurements were $A = 9.8 \times 10^{-2}$ *amps/cm² deg.²* and $b_0 = 8700$ ($\Phi_0 = 0.65$ *volts*).

(5) *Emission of positive ions from tungsten filaments in caesium vapor.* In connection with the further discussion of the phenomena of electron emission from caesiated tungsten, it is necessary to refer briefly to the important observations on the positive ion emission from tungsten filaments in caesium vapor.

In 1923 it was shown by Langmuir and Kingdon[76] that "A tungsten filament heated to 1200°K or more in saturated caesium vapor converts all caesium atoms which strike it into caesium ions. Thus when the filament is surrounded by a negatively charged cylinder, a positive ion current flows from the filament, which is independent of the filament temperature (above 1200°K) and independent of the applied potential, if this is sufficient to overcome the space charge effect of the positive ions. At lower voltages the currents follow the 3/2 power law, and are smaller than the corresponding electron currents obtainable from the same filament in the ratio of the square roots of the masses of the electrons and caesium ions.

"The reason that the caesium atoms lose their valency electrons so readily upon contact with the filament, is merely that the electron affinity of tungsten (Richardson work function) is 4.53 volts, while the electron affinity of a caesium atom (ionizing potential) is only 3.88 volts. Experiments showed, in fact, that if the work function for the filament is lowered to 2.69, by allowing a monatomic layer of thorium atoms to accumulate on the surface (by diffusion from the interior of a thoriated tungsten filament), the positive ion emission becomes negligible.

"The positive ions must be attracted to a tungsten surface because of the electron image force. Thus it is that below about 1200°K the caesium ions evaporate so slowly from a tungsten surface that this becomes partly covered by adsorbed caesium ions. The presence of these ions, however, lowers the electron affinity of the surface so that when about 20 per cent of the surface is covered, the work function falls below the ionizing potential of the caesium. With more caesium on the surface the caesium atoms which strike the filament no longer escape in the form of ions but remain in the atomic state. Thus the positive ion currents disappear below about 1100°K. But the lowering of the electron affinity raises the electron emission, and when the surface becomes more completely covered by caesium as the temperature is lowered, the electron emission rises to a maximum (of about 10^{-5} amps per cm^2 at 700°) and then decreases rapidly at still lower temperatures, in accordance with Richardson's equation."

Thus it is possible to determine the vapor pressure of caesium at any given

[75] L. R. Koller, Phys. Rev. **33**, 1082 (1929).

[76] I. Langmuir and K. H. Kingdon, Science **57**, 58 (1923); Proc. Roy. Soc., **A107**, 61 (1925).

bulb temperature from measurements of the positive ion current from a positively charged tungsten filament in presence of the vapor and maintained at a temperature above 1200°K. It was, in fact, by this method that the vapor pressure relation for caesium was determined which was used in a previous section. In the same manner T. J. Killian[74] has also measured the vapor pressure curve for rubidium and potassium.

The emission phenomena from different monatomic films on tungsten are well illustrated by Fig. 5 taken from a paper by I. Langmuir[77] in which log I is plotted against $1/T$. Thus it will be observed that at $T=1600°K$, the emissions from Th-W, pure W and O-W are in the ratio: $10^5:1:10^{-5}$. "The

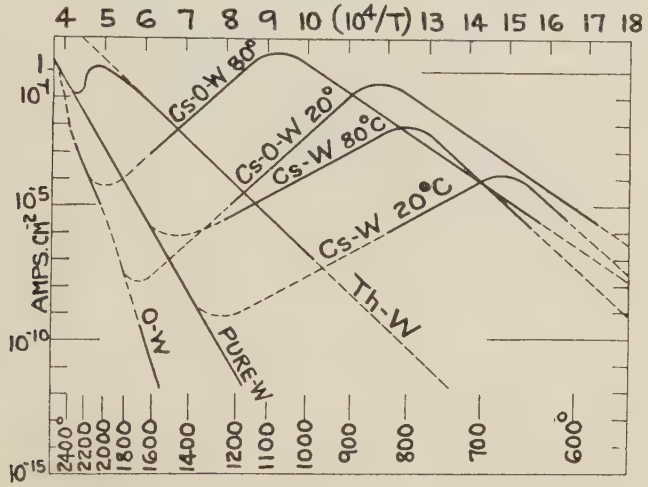


Fig. 5. Emission characteristics from various monatomic films on tungsten.

upper end of the line 'Th-W' shows a deviation from a straight line corresponding to the loss of adsorbed thorium by evaporation; at temperatures approaching 2400°K the emission becomes that of pure tungsten. . . . . The two curves marked 'Cs-W' give the electron emission of pure tungsten filaments in presence of caesium vapor saturated at 20° and 80°C. The falling off of emisson as the temperature is raised above 700° or 800°K is due to evaporation of caesium so that θ progressively decreases."

The curve marked O-W gives the emission from monatomic films of oxygen. "If heated to 1800°K or more, this film gradually evaporates off the filament as atomic oxygen (un-ionized), no tungsten oxide being found. When caesium vapor is brought into contact with a filament covered with an oxygen film, the filament gives an electron emission represented by the curves marked "Cs-O-W," which is almost 10^5 times greater than from "Cs-W" in the neighborhood of 1000°–1100°K."

(6) *Rate of activation of caesiated filaments.* Further observations on the rate at which tungsten and oxidized tungsten filaments become covered with adsorbed caesium atoms and on the values of θ for optimum electron emission

[77] I. Langmuir, Industrial and Eng. Chem. **22**, 390 (1930).

have been reported by J. A. Becker.[78] Fig. 6 taken from his paper shows typical curves for the rate of increase in emission after a Cs-O-W filament has been heated to 1040° and then changed suddenly to the value indicated on the curve. While at the higher temperature the positive ion current is measured, and from this result the rate is determined at which caesium atoms strike the surface. After reversing the voltage, to draw electrons from the filament, the temperature is changed suddenly to the lower value and the value of log I_e (the electron current) determined as a function of the time.

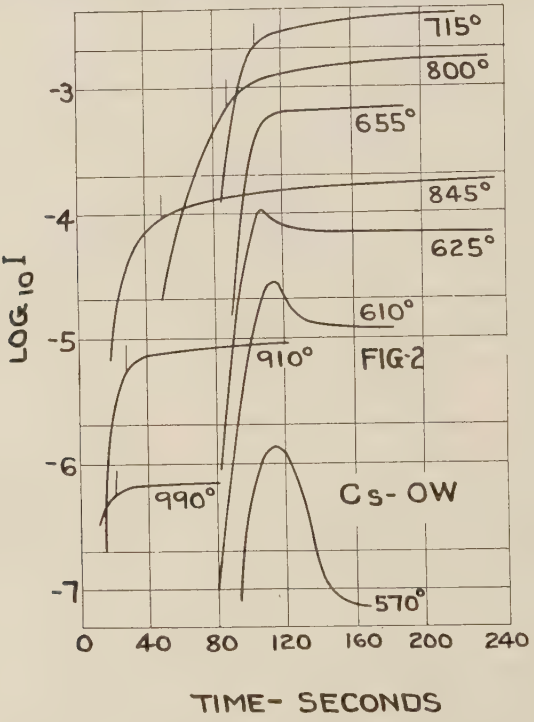


Fig. 6. Rate of increase in thermionic emission after a Cs-O-W filament has been heated to 1040°K and then changed suddenly to the value indicated on curve.

It will be observed that at temperatures below 650°K the emission passes through a maximum. "If the lower temperature is above 650°K the emission no longer passes through a maximum, but approaches its steady value rather more suddenly than one might expect. The arrows in Fig. 6 indicate the time necessary for sufficient caesium to have struck the surface to form an amount that is actually on the surface when equilibrium prevails. . . . . For Cs on W similar observations were made. The important point to note in these curves is that t_m, the time required to reach the maximum, is independent of T so long as $T < 650°K$. This suggests that below 650°K every atom that strikes the filament sticks to it until enough are adsorbed to give the optimum activity. . . . . This must mean that the fraction of the atoms

[78] J. A. Becker, Phys. Rev. **28**, 341 (1926).

that stick is the same at all these temperatures. It would be odd, indeed, if this fraction were anything but unity."

This conclusion was checked by the further observation that t_m varies inversely as i_p, the positive ion current per cm², that is, as the rate at which caesium strikes the surface.

From these observations, Becker concludes that the value of N_0, the number of Cs atoms per cm² for $\theta = 1$, is 3.7×10^{14} for Cs-W surfaces, and 4.1×10^{14} for Cs-O-W surfaces.

(7) *Equilibrium values of θ as a function of T and evaporation characteristics.* "For T between 1040° and 650° K the fraction of the surface covered, i.e., θ, should have values between zero and one. Hence the time required to reach the optimum activity should be less than t_m. This is actually found to be the case if we draw electrons, and from the times we can determine the values of θ for various filament temperatures." In this manner Becker has determined θ as a function of T for Cs-W and Cs-O-W.

Since at equilibrium the rate of arrival of caesium atoms is equal to the rate of evaporation, such data can be used to determine the evaporaton characteristics. If the filament is negative, only atoms evaporate, whereas if the filament is positive both atoms and positive ions evaporate. Becker finds that "the work necessary to remove an atom decreases as θ increases, while the work necessary to remove an ion increases with θ."

(8) *Work function for emission of electrons and positive ions.* From the above discussion it is evident that while in general the observations made by Becker are in agreement with those made by Langmuir and Kingdon, there are certain differences, While the latter assume that adsorbed surfaces cannot be more than one atom deep, and they conclude that maximum emission is obtained for $\theta < 1$, Becker concludes from his observations that at sufficiently low temperatures θ can exceed 1. From thermodynamical consideraations, for which the reader is referred to the original paper, Langmuir and Kingdon derive for Φ_e the work function for electron emission, the value 1.24 volts for $\theta = 0.90$ and conclude that Φ_e decreases linearly as θ increases, whereas Becker concludes that Φ_e has a minimum value of 1.36 volts for $\theta = 1$ and shows a non-linear increase for $\theta < 1$, until it reaches the value 4.52 for pure tungsten. For the heat of evaporation of positive ions of caesium from pure tungsten, Langmuir and Kingdon derive the value 4.0 volts for $\theta = 0$, which increases to 4.3 volts for $\theta = 0.9$. In the case of caesium films on adsorbed oxygen, the heat of evaporation of ions is deduced as 5.13 volts for $\theta = 0.95$.

(9) *Adsorption phenomena.* It is evident from the preceding discussion that the emission phenomena observed for films of caesium (and presumably other alkali metals) on tungsten or tungsten covered with a monatomic film of oxygen, are extremely complex. The photoelectric behavior of caesium on oxidized copper and silver shows similar characteristics and undoubtedly in both the thermionic and photoelectric emission, the work function varies considerably with variations in the nature of the layer on which the alkali metal is deposited. As shown by Langmuir and Kingdon, these phenomena are to be interpreted as due to variations in Φ and are consequently intimate-

ly related to the adsorption characteristics of caesium on various surfaces. More recently J. A. Becker has investigated these phenomena in greater detail[79] than in his first paper and has determined the rate of evaporation of caesium from Cs-O-W surfaces as a function of θ and T. The results obtained are shown graphically in Fig. 7. From these observations it is possible to calculate the mean life, τ, of a caesium atom as a function of θ and T. Thus at $T = 700°K$, τ decreases from 665 seconds for $\theta = 0.8$ to 28 seconds for $\theta = 1.00$, and with increase in temperature for constant value of θ, τ decreases.

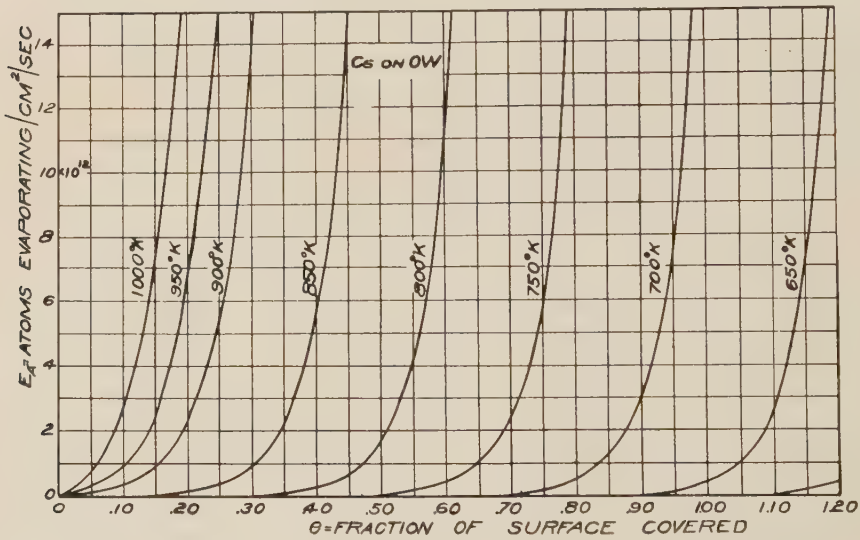


Fig. 7. Rate of evaporation of caesium from Cs-O-W surfaces as a function of temperature and of fraction of surface covered, θ.

IV. Calorimetric determination of work function for metals

As shown by Richardson the emission of electrons involves a cooling effect which may be regarded as the latent heat of evaporation of electrons (either at constant volume or constant pressure). From measurements on this cooling effect, it is therefore possible to calculate a value of the work function corresponding to the constant b (or b_0) in the emission- temperature equations (1) and (2).

The theory underlying such measurement as well as a description of the experimental arrangement, are given in the treatises of Richardson and of Schottky and Rothe. Also the publications of C. Davisson and L. H. Germer[80] give very complete details.

The cooling effect as determined experimentally may be designated by the volt equivalent value Φ_c. From this value the constants b and b_0, or Φ and Φ_0 respectively, are determined by the relations

[79] J. A. Becker, Trans. Am. Electrochem. Soc. May, 1929.

[80] C. Davisson and L. H. Germer, Phys. Rev. **20**, 300 (1922); **24**, 666 (1924).

$$\frac{b_0 k}{e} = \Phi_0 = \Phi_c - \frac{2kT}{e}$$
$$\frac{bk}{e} = \Phi = \Phi_0 + \frac{3}{2}\frac{kT}{e} = \Phi_c - \frac{kT}{2e} \qquad (21)$$

where T is the temperature of the cathode in the calorimetric determination.

"The first experiments to detect and measure this effect were made by Wehnelt and Jentzsch[81] using the emission from lime-coated wires." While their results were in qualitative agreement with values from emission measurements, the first definite proof of the phenomenon was obtained by H. L.

TABLE XII. *Calorimetric determinations of work function for homogeneous cathodes.*

Author	Material	T	Φ (cal.)	Φ_0 (therm.)
Cooke and Richardson[83]	Os		4.7	—
	W	1922	4.24	4.48
H. Lester[84]	W		4.48	4.48
Davisson and Germer[80]	W	2270	4.52	4.48
Michel and Spanner[85]	W	2420	4.57	5.0
Wehnelt and Liebreich[86]	Pt		5.9	5.0
H. Lester[84]	Mo		4.59	4.33
	Ta		4.51	4.2
	C		4.55	—

Cooke and Richardson from experiments with osmium filaments. Subsequently similar measurements were carried out with other metals by the same investigators and others. Table XII[82] gives the results of these investigations under Φ_0(cal), and for comparison are given the values of Φ_0 from emission data on similar filaments. Davisson and Germer and also Lester made determinations by both methods on the same filaments.

The converse phenomenon, that is, the absorption of heat by metals when electrons impinge on them, has also been observed by Richardson and Cooke[87] and these investigators have determined, in this manner, work functions for a number of metals. However, the results obtained cannot be regarded as very accurate.

More recently Van Voorhis[88] has measured the heat of condensation of electrons by a method involving the use of Langmuir collecting electrodes[89] in a gas discharge. The values of Φ_0 found by this method for molybdenum

[81] Wehnelt and Jentzsch, Verh. d. Deutsch. physik. Ges. **10,** 610 (1908); Ann. d. Physik **28,** 537 (1909).

[82] From treatise by Schottky, Rothe and Simon, p. 136 with slight additions. See also Richardson's book p. 188.

[83] Cooke and Richardson, Phil. Mag. **25,** 624 (1913); **26,** 472 (1913).

[84] H. Lester, Phil. Mag. **31,** 197 (1916).

[85] Michel and Spanner, Zeits. f. Phys. **35,** 395 (1925).

[86] Wehnelt and Liebreich, Physikal. Zeits. **15,** 548 (1914).

[87] Richardson's book, pp. 189–195.

[88] Van Voorhis, Phys. Rev. **30,** 318 (1928); Compton and Van Voorhis, Proc. Nat. Acad. Sci. **13,** 336 (1927).

[89] Langmuir, G. E. Review **26,** 731 (1923); Langmuir and Mott Smith, G. E. Review **27,** 449, 538, 616, 762, 810 (1924).

were as follows: 4.76 volts in argon, 4.04 or 4.35 volts in hydrogen, and 4.77 or 5.01 volts in nitrogen. The double values were due to differences in treatments of the surface. It will be observed that these values are higher than those obtained thermionically or by calorimetric methods in vacuum. Likewise, Schottky and von Issendorff[90] have deduced, from energy measurements on probe electrodes in a mercury discharge, that the work function in case of iron and nickel is between 4 and 5 volts. Whether this method can be developed to yield accurate data on Φ_0 seems questionable in view of the many corrections which have to be introduced in order to determine the desired result.

V. Electron Emission from Metals at the Melting Point

Wehnelt and Seeliger[91] measured the emission (of both electrons and positive ions) from copper and silver at the melting point. Their results for a and Φ (corresponding to b in equation (1)) were as follows:

Metal	a (amps/cm^2 deg.$^{1/2}$)	Φ
Cu—solid	9.22×10^5	3.85 volt
fluid	4.93×10^5	3.71
Ag—solid	3.41×10^3	3.09
fluid	1.77×10^3	2.9

The most careful and extensive investigations on this subject have been carried out by A. Goetz[92] since 1923. A specially designed vacuum furnace was used and the thermionic emission was observed for a series of temperatures below and above the melting points of the metals.

In the first publication, plots of $\log I/T$ vs. $1/T$ are given for copper, iron, and manganese. In the case of iron it was observed that at the transition temperatures for the different crystal modifications, the value of a in Richardson's equation changed, but not that of b. At the melting points of all three metals the value of b was observed to increase considerably.

Fig. 8 shows a typical plot of $\log (i/T^2)$ vs. $1/T$ as obtained for copper (melting point 1358°K) in the more recent investigation. From these observations, it is found that A and Φ_0 vary with T as shown in Table XIII.

Table XIII. *Emission constants for copper* (Goetz).

T	$b_0\times10^{-4}$	Φ_0	A
1316	5.02	4.33	65
1380	1.98	1.79	6.0×10^{-10}
1400	3.19	2.75	4.0×10^{-5}
1430	4.55	3.92	6.2×10^{-1}
1480	6.37	5.50	1.5×10^5
1530	7.75	6.69	1.2×10^9

[90] Schottky and von Issendorff, Zeits. f. Physik **26,** 85 (1925).

[91] Wehnelt and Seeliger, Zeits. f. Physik **38,** 443 (1926).

[92] A. Goetz, Phys. Zeits. **24,** 377 (1923), **26,** 206 (1925); Zeits. f. Physik **42,** 329 (1927); **43,** 531 (1927). Schottky and Rothe also give a very complete summary of these investigations in their treatise.

Fig. 9 shows a plot of Φ_0 vs. T. At the melting point the value decreases abruptly and then increases gradually. Similar results were observed for silver and gold. The values of the emission constants for these metals near the melting point are shown in Table XIV.

TABLE XIV. *Emission constants for silver and gold* (Goetz).

Metal	State	T	$b_0 \times 10^{-4}$	Φ_0	A
Ag	solid	1233	4.64	4.0	
	fused	1240	1.98	1.71	3.0×10^{-8}
	fused	1450	3.78	3.26	2.9×10^{-2}
Au	solid	1315	5.00	4.32	40
	fused	1385	2.91	2.51	5.8×10^{-6}
	fused	1530	4.98	4.30	13

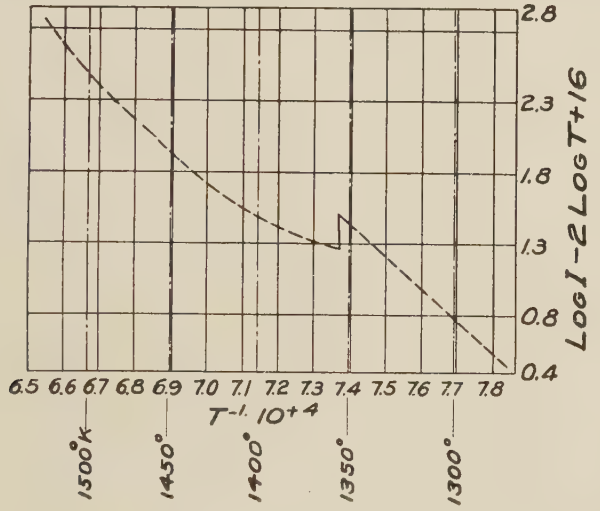


Fig. 8. Variation of thermionic emission through the melting point of copper.

As Schottky has pointed out, it is difficult to attach any interpretation to these results. They indicate, however, that the work function varies considerably with the temperature.

The emission from solid and fused potassium and sodium has been investigated by A. F. A. Young[93] and Young and Richardson.[94] Earlier investigations on these metals were undoubtedly inaccurate because of the presence of gas. Young's observation in the case of potassium did not yield any constant values of the emission constants, since they varied with both anode voltage and temperature. Thus the value of $b \times 10^{-4}$ varied from 1.58 to 1.32 at 200°C, and from 0.576 to 0.505 at 30°C, with corresponding variations in a. However, no abrupt change in work function was observed at the melting point.

[93] A. F. A. Young, Proc. Roy. Soc. **104A**, 611 (1923).

[94] Young and Richardson, Proc. Roy. Soc. **107A**, 377 (1925).

The measurements made by Young and Richardson gave values of $b \times 10^{-4}$ varying from 1.90 to 0.693. It is of interest to note that a plot of values of $\log A$ vs. b_0 shows an approximately linear relation. In order to interpret the observations it was suggested that two modifications of potassium are present, the more active one concentrating at the surface in the form of "patches."

VI. Oxide Coated Cathodes

The emission from oxides of the alkaline earth metals was first observed by. A. Wehnelt[95] and has since then been the subject of many investigations not only on account of its technical importance, but also because of the complex nature of the phenomena which characterize the behavior of these materials when used as coatings on metallic filaments. A bibliography of the

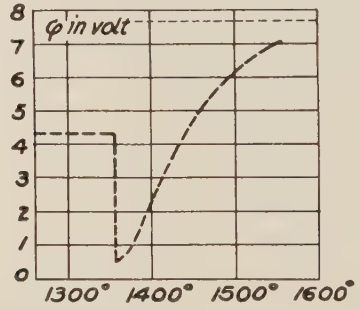


Fig. 9. Variation of work function of copper through the melting point.

earlier work (to 1920) in the physics and chemistry of oxide coated cathodes has been given by H. D. Arnold[96] and a discussion of this work and of the various theories theories held before that time on the nature of the emission phenomena will be found in that paper as well as in the treatises by Richardson and Schottky, Rothe and Simon.[97]

In reviewing the numerous publications which have appeared on this subject, it has been considered best to discuss this work under the following headings:

(A) Preparation of cathodes.
(B) Determination of temperature.
(C) Emission constants.
(D) Determination of work function calorimetrically.
(E) Activation and de-activation phenomena and theory of emission.

A. Preparation of oxide coated cathodes

The various methods for producing filaments coated with alkaline earth oxides are described in the treatise by Schottky, Rothe and Simon[98] and

[95] A. Wehnelt, Ann. d. Physik **14**, 425 (1904).
[96] H. D. Arnold, Phys. Rev. **16**, 70 (1920).
[97] Schottky, Rothe and Simon, vol. 13, of Wien Harms "Handbuch der Experimentalphysik."
[98] See C. Davisson, I. C. T. **6**, 53 (1929).

also by H. D. Arnold,[96] L. R. Koller,[99] E. F. Lowry[100] and E. R. Wagner[101]. As core material, platinum (alloyed with six to ten percent iridium, or with ten percent rhodium) or with 5 percent nickel, or an alloy of nickel, cobalt, iron and titanium designated as "Konel" metal[100] may be used. The oxides have also been used successfully on a tungsten filament which has been nickel plated and then oxidized to form a film of oxide on the surface. In many cases the metals are used in the form of narrow ribbons or twisted wires.

The oxides (usually a mixture of equal parts by weight of BaO and SrO) are put on the core metal either in the form of carbonates or nitrates. According to Arnold, "Barium in the form of carbonate, and strontium in the form of hydroxide or carbonate, was mixed with some carrier such as resin or paraffin, which would burn away when heated in the air. In the coating process four applications of the strontium mixture were followed by four of the barium mixture, and this process was then repeated, making a total of sixteen separate applications. After each application the wire was raised momentarily to a temperature of about a thousand degrees, which burned away most of the organic carrier. When the coating was complete the wire was heated to about 1200° for two hours. At the end of this time there remains a fairly heavy coat of BaO and SrO (from 2 to 3 milligrams per sq. cm surface), while next to the core is a firmly adhering layer which is built up due to chemical reactions between the coating and the core. Analysis shows this coating to consist of barium and strontium combined with platinum, rhodium and iridium, the compound present in largest amount being barium platinate ($BaPtO_3$). The compound with rhodium seems to be more readily formed, but due to the small percentage of rhodium present this compound makes up only a small fraction of the total.

"The filament thus formed can be handled without undue precaution as long as it is not exposed to moisture or carbon dioxide."

In the case of the nitrates, according to Koller, "The method of applying the coating was to dip the wire into a solution containing respectively three percent of barium nitrate and two percent of strontium nitrate and then burn it for a few seconds at a high temperature in an atmosphere of carbon dioxide. This process was then repeated until the coating reached the desired thickness. About 40 cycles were usually sufficient.

"Such a coating consists largely of the carbonates, as can readily be demonstrated by dipping a filament in hydrochloric acid, which results in a rapid evolution of gas. This evolution of gas takes place even after a filament has been burned for several hundred hours."

The coating can also be applied in the form of a water paste consisting of equal parts of the carbonates of barium and strontium. This is then baked on the filaments in an atmosphere of CO_2 (Lowry). Or the oxides may be used in the form of a suspension either in alcoholic solution of colophonium or pure paraffin oil. The mixture is then applied to the wire in thin layers

[99] L. R. Koller, Phys. Rev. **22** 671 (1925).
[100] E. F. Lowry, Phys. Rev. **35**, 1367 (1930).
[101] E. R. Wagner, Electronics, July (1930), p. 178.

followed by heating in air first for a short period to 500°C and then to 1000°C. (W. Espe[102].)

In general it is not advisable to heat the filaments to too high a temperature before they are inserted in the tube and de-gassed on the pump. Such cathodes are apt to be very difficult to activate.

The subsequent treatment of the cathodes, during exhaust, is necessary in order to form or activate them. The first stage consists in glowing the filaments to a moderately high temperature to decompose the carbonates or nitrates, and eliminate occluded gases. The anodes and other metal parts should be treated after this operation with high frequency, to de-gas them. The second stage consists in the activation or formation process. J. A. Becker[103] designates this also as the "breakdown process," and describes it as follows :"This process consists in glowing the oxide coated platinum filament at about 1500°K for several minutes. A rather high positive potential gradient is then applied to the filament for something like two to thirty minutes. During this time the emission from the filament increases greatly. When it attains rather large values such as 0.1 to 1.0 amps. per cm^2, the applied field and the filament temperature are so adjusted that the current is just limited by space charge." In the case of nickel or Konel filaments the maximum temperature should not be over about 1000°C. Approximately a similar procedure has been used by W. Espe[102] in his investigations on oxide coated filaments.

The activation may also be accomplished by thermal treatment alone without the application of anode voltage. This seems to apply specially to filaments of nickel or Konel. In the case of the latter Lowry has shown that a treatment at 8.5 watts per cm^2 (corresponding to about 1075°C) for several minutes, followed by a more prolonged treatment at about 950°C is sufficient to activate the cathodes. Lowry also states in the same connection that, "Similar results have been obtained with oxide coated platinum filaments." The experience in the Research Laboratory at Schenectady has shown that application of anode voltage is not essential, although in some cases especially in those in which platinum or tungsten filaments are used as core material, the operation of drawing electrons from the cathode often increases the rate of activation. Other investigators have also observed activation of oxide coated platinum filaments by thermal treatment exclusively (M. S. Glass,[104] Reimann and Murgoci[105] and Koller[99]).

Mention should also be made in this connection of a type of oxide coated cathode developed in Europe and used to a considerable extent in vacuum devices made on the continent.[97] A solution of barium azide (BaN_6) is brushed on the inside of the metal (usually nickel) anode and the cathode is made of tungsten which has been electroplated with an extremely thin film of copper or nickel and then oxidized by heating in air. During exhaust, the

[102] W. Espe, Wiss. Veroeff. d. Siem. Konz. **5**, 29, 46 (1927).
[103] J. A. Becker, Phys. Rev. **34**, 1323 (1929).
[104] M. S. Glass, Phys. Rev. **28**, 521 (1926).
[105] A. L. Reimann and R. Murgoci, Phil. Mag. **8**, 440 (1930).

anode is heated by the high frequency method, the azide dissociates and metallic barium is deposited on the oxide layer on the cathode. The latter is then glowed for a considerable time to activate the deposit and the tube is then sealed off. This method apparently offers two advantages: firstly, much smaller filament diameters of wire may be used, which therefore require smaller heating currents, and secondly, the barium volatilized all over the bulb walls acts as a "getter" and therefore maintains an even better vacuum in the tube than can be obtained with the more commonly used magnesium getter. However, in this country, this type of cathode has not been commercially utilized to any extent, as far as the writer is aware.

The various methods used in the manufacture of oxide coated cathodes have also been described by Hodgson, Harley and Pratt,[106], and a comparison given in their paper of the emission from the different filaments during life.

B. Temperature scale for oxide coated cathodes.

As might be inferred from the description of the process of preparing and activating this type of cathode, the thermal emissivity varies considerably with variations in materials and in method of forming.

For the filaments used by the Western Electric Co. in their vacuum tubes, Dr. C. Davisson kindly furnished the writer the following information:[98]

Material of filament core: Platinum 95 percent, nickel 5 percent.

Electrical resistivity:

$$\rho = 2.2 \times 10^{-5}(1 + 0.00208t - 4.6 \times 10^{-7}t^2) \tag{22}$$

ohms cm, where t = Centigrade temperature.

The coating is a mixture of the oxides of barium and strontium plus a small admixture of nickel (and platinum) which is transferred from core to coating during the preparation and activation of the filament.

Emissive Power:

$$\epsilon = 0.4 + 2.5 \times 10^{-4}T \text{ for the range } 800° - 1200°\text{K.}''$$

From the last relation it follows that the total energy radiated is given by the relation,

$$W = 5.735 \times 10^{-12}(0.4 + 2.5 \times 10^{-4}T)T^4 \text{ watts/cm}^2 \tag{23}$$

$$= 3.727 \text{ watts/cm}^2 \text{ at } T = 1000° \text{ K.}$$

For *pure platinum* C. Davisson and L. H. Germer[107] give the following relation for the resistivity:

$$\rho_t = 0.4217(1 + 0.003946t - 5.85 \times 10^{-7}t^2) \tag{24}$$

where t is the temperature in degrees Centigrade.

F. Detels[108] obtained from a series of measurements with blank and oxide-coated platinum wires (ten percent iridium) a linear relation for the resistivity of the form

[106] B. Hodgson, L. S. Harley and O. S. Pratt, Journal I. E. E. **67**, 762 (1929).

[107] C. Davisson and L. H. Germer, Phys. Rev. **24**, 666 (1924).

$$\rho_t = \rho_0(1 + 2.65 \times 10^{-4}t) \tag{25}$$

where t = Centigrade temperature.

For coated tungsten wires the ratio of resistance at the high temperature to that at room temperature also furnishes a convenient method of determining the temperature, provided of course, that a proper correction is made for cooling effect of leads.

E. F. Lowry[100] has given brightness temperature measurements for ribbon filaments of coated platinum and Konel. The latter is a much better radiator. By enlarging the curve shown in the original paper, the values given in Table XV were obtained for the energy radiated as a function of T. The second column gives the values derived from Davisson's relation, equation (23), and the third column, Lowry's values for Konel metal. The last two columns give the hot to cold ratio of resistance for Western Electric and platinum-iridium (10 percent) alloy, according to equations (22) and (25) respectively.

TABLE XV. *Resistivity and energy radiated for coated filaments (BaO+SrO)*

T	Energy radiated (watts/cm²) W.E. Pt-Ni	Konel	ρ_t/ρ_0 W.E. Pt-Ni	Pt-Ir
300°K	0.022		1.00	1.00
700	0.79	1.66	1.80	1.11
800	1.41	3.01	1.96	1.13
900	2.35	4.72	2.12	1.16
1000	3.73	6.57	2.26	1.19
1100	5.67	8.91	2.40	1.21
1200	10.70		2.53	1.24
1300	15.15		2.65	1.27
1400	20.93		2.75	1.29
1500	28.31		2.85	1.32

While the temperature-coefficient of resistance is much smaller than that of the energy radiated, it has the considerable advantage that the value is the same for both coated and blank wires. Therefore, if it can be measured accurately for the latter, the temperature determined for the coated wire is independent of any knowledge of the thermal emissivity, which in the case of coated wires may vary even for filaments apparently identical in composition and treatment. On the other hand, as mentioned previously, the error due to cooling effect of leads may be considerable on account of both the short lengths ordinarily used and the relatively low temperatures at which the filaments are operated

Because of these difficulties in the accurate determination of temperature, it has been found more convenient often to compare the electron emission from various coated filaments at the same watts imput. (Such a series of plots is shown by Arnold[96] and also by R. W. King[109]. This topic, however, will be discussed further in the following section.)

[108] F. Detels, Jahrb. d. drathlos. Telegr. **30**, 10, 52 (1927).
[109] R. W. King, Bell System Tech. J. **2**, No. 4 (1923).

C. Thermionic emission constants for coated filaments.

As pointed out by H. Rothe[110] and W. Espe,[102] the electron emission from coated filaments does not exhibit the same distinct saturation values as those observed for homogeneous metal filaments. The space current increases with voltage very rapidly, especially at higher temperatures, even when it is considerably below the limit put by space charge relations. Furthermore the

TABLE XVI. *Thermionic emission constants for oxide-coated filaments.*

Year	Oxide	a	$b\times10^{-4}$	ϕ (volts)	A	$b_0\times10^{-4}$	Φ_0	Remarks	Ref
1904	BaO	1.11×10^{8}	4.5	3.9			3.65	Pt. core	(95)R
	CaO	7.2×10^{7}	4.3	3.75			3.48	" "	(95)R
1907	CaO	6.36×10^{11}	4.8	4.15			3.9	" "	(111)R
1908	CaO	1.75×10^{7}	4.3	3.75	4.2×10^{-8}	4.05	3.48	" "	(112)R
1908	Al_2O_3	6.40	3.73	3.21	2.1×10^{-5}	3.48	3.00	" "	(113)R
	BaO	4.70×10^{7}	4.16	3.58	2.7×10^{2}	3.99	3.44		
	BeO	1.03	2.39	2.06	4.56×10^{-6}	2.19	1.88		
	CaO	4.30×10^{7}	4.03	3.48	2.49×10^{2}	3.86	3.33		
	CdO	0.37	3.02	2.60	1.65×10^{-6}	2.82	2.43		
	CeO_2	1.95×10^{3}	3.71	3.20	8.62×10^{-2}	3.51	3.02		
	Co_2O_3	6.67×10^{3}	4.97	4.28	2.17×10^{-2}	4.72	4.06		
	CuO	3.5×10^{-3}	2.25	1.94	1.55×10^{-8}	2.05	1.76		
	Fe_2O_3	3.57×10^{3}	4.69	4.04	1.16×10^{-2}	4.44	3.82		
	La_2O_3	6.9×10^{2}	3.79	3.26	3.03×10^{-3}	3.59	3.10		
	MgO	3.37	3.95	3.40	1.10×10^{-5}	3.70	3.19		
	NiO	2.79×10^{4}	5.12	4.41	9.1×10^{-2}	4.87	4.19		
	SrO	5.07×10^{7}	4.49	3.87	2.58×10^{2}	4.30	3.71		
	ThO_2	35.	3.56	3.06	1.54×10^{-4}	3.36	2.89		
	Yt_2O_3	1.86×10^{4}	3.63	3.13	8.22×10^{-2}	3.43	2.95		
	ZnO	0.31	3.51	3.02	1.0×10^{-6}	3.26	2.81		
	ZrO_2	6.57×10^{3}	3.66	3.15	2.9×10^{-2}	3.46	2.98		
1914	Nernst Fil.	1.11×10^{6}	4.6	4.0					R
1916	CaO			2.54					(114)
1920	BaO+SrO	8×10^{4} to 2.4×10^{5}	1.94 to 2.38	1.55 to 1.90	.418 to 1.23	1.76 to 2.20	1.51 to 1.89	Pt. alloy core	(96)
1924	Al_2O_3	3.3×10^{6}	4.58	3.95	1.4	4.37	3.77	Pt.-alloy core	(115)
	B_2O_3	1.2×10^{5}	4.81	4.15	0.49	4.59	3.95		
	BaO	$8.\times10^{5}$	2.15	1.85	4.0	1.97	1.70		
	BeO	9.8×10^{5}	4.01	3.45	4.9	3.83	3.30		
	CaO	4.8×10^{6}	2.78	2.4	25.7	2.60	2.24		
	MgO	8.5×10^{6}	3.49	3.01	45.7	3.31	2.85		
	SrO	1.3×10^{6}	2.49	2.15	6.9	2.31	1.99		
	Sc_2O_3	2.5×10^{6}	4.18	3.6	10.7	3.97	3.42		
	SiO_2	1.6×10^{6}	5.57	4.8	5.0	5.31	4.58		
	ThO_2	4.6×10^{3}	3.65	3.15	0.016	3.41	2.94		
	TiO_2	4×10^{4}	4.64	4.0	0.13	4.39	3.78		
	Yt_2O_3	1.26×10^{6}	3.60	3.1	5.4	3.39	2.92		
	ZrO_2	$1.\times10^{5}$	4.18	3.6	0.35	3.94	3.40		
1925	BaO+SrO				1.07×10^{-3}	1.21	1.05	Pt-Ir core	(99)
1926	BaO+SrO	100	1.073	0.92	9.6×10^{-4}	0.953	0.82	German tube	(110)
		60	.969	.83	5.7×10^{-4}	.849	.73	" "	
		10	.747	0.64	9.8×10^{-5}	.627	.54	" "	
		1.66×10^{3}	1.309	1.13	1.6×10^{-2}	1.189	1.02	" "	
		2.5×10^{4}	1.444	1.24	0.24	1.324	1.14	Dutch "	
1927	SrO	0.75×10^{3}	1.64	1.41				200 V	(102)
		2.4×10^{3}		1.44				300V	
	CaO+BaO+SrO				8.3×10^{-3}	1.44	1.24	See table XVII	(102)
1927	BaO+SrO				2.97×10^{-2}	1.58	1.36		(108)
					9.3×10^{-4}	1.58	1.36		
1928	BaO				2.5		1.66	W core	(116)
1928	BaO+SrO				1.0×10^{-2}	1.16	1.00	West Electric (Zero Field) See table XVIII.	(97)

emission is apt to change with time and there are pronounced fatigue effects, which may lead to erroneous values of the emission constants. In fact, as will be pointed out more fully in the following section, the behavior of oxide

[110] H. Rothe, Zeits. f. Physik **36,** 737 (1926).
[111] F. Horton, Phil. Trans. **A207,** 149 (1907).
[112] F. Deininger, Ann. d. Physik **25,** 258 (1908).
[113] F. Jentzsch, Ann. d. Physik **27,** 129 (1908).
[114] W. Germershausen, Phys. Zeits. **16,** 104 (1915); Ann. d. Physik **51,** 705 (1916).
[115] H. J. Spanner, Ann. d. Physik **75,** 609 (1924).
[116] J. M. Eglin, Phys. Rev. **31,** 1127 (1928).

coated cathodes is so different from that of filaments of pure metals that it is extremely difficult to obtain precision data on emission constants comparable with those given in Table IV. This is especially true of the values of A.

Table XVI gives emission constants obtained by the various investigators. The relative accuracy varies with the conditions under which the observations were made, and an attempt will be made in the following paragraphs to point out those data which are to be regarded as more reliable.

In the case of most of the data, expecially those obtained by earlier investigators, no corrections were made apparently for either temperature distribution along the cathode or the effect of field strength. Since oxide coated filaments are operated at comparatively low temperatures (below 1200°K), the lead loss correction is extremely important. Moreover, since the resistitivity of platinum or nickel is less than that of tungsten, the voltage drop along an oxide coated filament is also quite low. (This is specially true of the filaments used in commercial vacuum tubes.) Hence the errors due to neglecting these considerations are apt to be large.

With regard to the effect of field strength it is certain from the work of J. A. Becker and others that for the type of filament under consideration, Schottky's relation is not valid. Consequently, either the extrapolation to zero field strength must be made on the basis of other considerations, or emission data must be investigated to as low field strengths as possible.

For these two reasons, but mainly on account of ignorance of the proper method of correcting for effect of field strength, the values of the thermionic constants given in Table XVI cannot be regarded, in the large majority of cases, as more than a first approximation to the true values of A and b_0 (or Φ_0). Undoubtedly further investigations in this field are necessary before such precise data will be available.

Of the earlier data, those published by Jentzsch[113] are of interest because of the number of oxides investigated. A comparison of these data with those obtained by Spanner[115] in 1924 shows marked differences in a number of cases. The paper by the latter gives the values of Φ (not Φ_0) and plots of log. $(I/T^{1/2})$ vs. $1/T$. These plots were enlarged and used to determine values of a. From these, the values of A and b_0 were calculated in the manner described in connection with Table IV. Spanner used short filaments and therefore his data probably involve large lead loss corrections. However, since the values calculated from his plots for W, Ta, and Pt are not radically different from those obtained by other investigators, it seems reasonable to conclude that, on the whole, his data are probably fairly accurate.

From his measurements Spanner concluded that Φ, the work function corresponding to b in equation (1), is a simple function of the atomic number N, and of the number of valence electrons, Z, for the metal, of the form:

$$\Phi = \frac{7Z^{3/2}}{4N^{1/2}} + 1 \text{ volt}.$$

It is, however, questionable whether this relation has any significance.

The emission constants reported by Arnold[96] in 1920 are the first satis-

factory data, but are not as accurate as those reported on similar filaments by Davisson[98] in 1928. Because it is of great technical importance to obtain electrons as "cheaply" as possible, the investigators from the Bell Laboratories have reported their observations on thermionic emission in the form of plots such as those shown in Fig. 10, taken from a paper by R. W. King.[109] The emission in milliamps/cm² is plotted against the power input (watts/cm²)

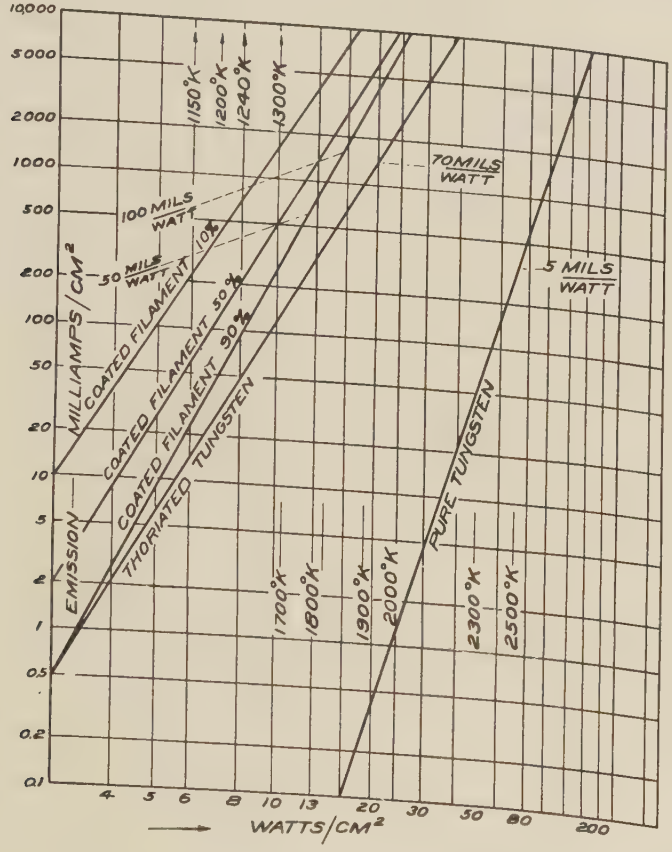


Fig. 10. Emission from coated filaments as a function of power input.

on specially designed cross-section paper, as explained by Arnold.[96] The figure shows the variation in emissions for commercially manufactured filaments. Plots for thoriated tungsten and pure tungsten are shown for comparison.

The remarkably low value for Φ_0 obtained by Koller[99] after careful evacuation and prolonged activation has also been observed by other investigators subsequently.

The data reported by Rothe[110] were taken on commercial tubes, four of German manufacture and one of Dutch. No correction was made for cooling effect of leads or field strength. Rothe himself considers the values of A as not very reliable. A noteworthy feature of these observations is the relatively

low temperature range over which it was found necessary to operate the filaments. Thus, on the tube from Holland, the emission varied from 1.2×10^{-5} amp./cm$_2$ at $T=595°$K to 76.5×10^{-3} at $T=905°$K. In calculating values of A and b_0 from values of a and b given in the original paper, the average temperature was therefore taken as $T=800°$K. At $T=1000°$K the emission calculated from Rothe's values of a and b varies from 69 to 800 milliamps/cm^2 (see Table XIX). These values are higher than those observed by most of the other investigators, except under special conditions. This may be accounted for either by assuming that Rothe's values for T are too low or else that he was using filaments made by the barium azide process. While nothing is stated by him regarding the manner of preparation of the cathodes, it is quite probable that these were of the azide type, since at that time a considerable number of tubes were made by this process in Europe and especially in Holland. It should, however, be stated that in the paper by Becker[103] emission data are mentioned which are of the same magnitude as those observed by Rothe.

TABLE XVII. *Work function for oxide coated filaments (W. Espe[102]).*

Coating	$b\times10^{-4}$	Φ	$b_0\times10^{-4}$	Φ_0	Anode volts
CaO—1	1.98	1.71	—	—	100
—2	1.97	1.70	—	—	200
—3	2.22	1.90	2.02	1.74	200
—4	2.27	1.96	2.04	1.76	200
—5	2.23	1.92	2.08	1.79	300
SrO—1	1.67	1.44	1.44	1.24	300
—2	1.635	1.41	1.46	1.26	200
—3	1.67	1.44	1.51	1.29	300
—4	1.68	1.45	1.50	1.29	200
BaO—1	1.27	1.095	1.13	0.97	300
—2	1.265	1.09	1.13	0.97	300
—3	1.29	1.11	1.16	1.00	300
—4	1.31	1.13	1.19	1.03	300
CaO	2.24	1.93	2.05	1.77	
SrO	1.66	1.43	1.475	1.27	
BaO	1.29	1.11	1.15	0.99	

Results of comprehensive investigations in this field have been published by W. Espe[102] and F. Detels.[108] The former has emphasized the effect of the anode voltage at which the emission is observed. Representative results for SrO on Pt-Ir are given in Table XVI. For coatings of CaO to which 5 percent of BaO and SrO were added he observed values of A ranging from 0.2×10^{-3} to 20×10^{-3} amps./cm^2 deg^2, according to degree of activation. The values of the work functions obtained by him for each of the oxides are given in Table XVII, taken from his paper. The values for CaO-1 were calculated from preliminary tests, those for CaO-2 were obtained with a "heavy" layer, While all the other values were obtained with "thin" layers. The final values deduced by Espe as most reliable are given in the last three lines in the table. The maximum error in b (or b_0) is stated to be about 300.

Detels[108] observed that during activation both A and b_0 decrease simul-

taneously. These observations will be discussed in a subsequent section. The values of emission constants given in Table XVI are typical of minimum values obtained after a prolonged period of activation. Corrections were introduced for temperature distribution along the filament. In fact, the original paper gives quite a comprehensive treatment of this topic. Since the emission-voltage curves apparently flattened out at 100–200 volts anode potential, the saturation values were used to calculate A and b_0. Therefore no correction was introduced for effect of field strength.

TABLE XVIII. *Most probable characteristics of Western Electric filaments (C. Davisson[97]).*

Temp:—	900°K	950°K	1000°K	1050°K	1100°K
I_0 (milliamps/cm²)	20	45	90	170	310
P_r (power radiated, watts/cm²)	2.3	3.0	3.7	4.6	5.6
P_e (power absorbed by emission. watts/cm²)	0.02	.045	.09	·17	.31
Efficiency (milliamps/watt) $I_0/(P_r+P_e)$	8.6	15	24	31	53
Life (in thousands of hours)	(730)*	(170)*	20	20	7.4

* These figures are computed.

Eglin's observations[116] are of interest because they represent results obtained for films of metallic barium deposited on tungsten in vacuo, by evaporation from regular coated filaments.

The values forWestern Electric filaments were furnished by Dr. Davisson[97] and represent the average of a very large number of observations. It will be observed that the value of Φ_0 is stated to be that obtained for zero field strength. Table XVIII, also from the same source, gives the most probable characteristics of this type of cathode at anode potentials comparable

TABLE XIX. *Comparison of emission from BaO+SrO at T=1000°K.*

Author	Year	b_0	I(milliamps/cm²)
Arnold	1920	1.76	9.5
		2.20	.35
Spanner	1924	1.97	11.25 (BaO)
Koller	1925	1.21	6.6
Rothe	1926	.95	69.
		.85	117.5
		.63	182.
		1.19	110
		1.32	800
Espe	1927	1.44	4.6
Detels	1927	1.58	44.1
			1.4
Davisson	1928	1.00	91.8

with 150 volts. With regard to life, (see last line of Table XVIII), Dr. Davisson states that, "For filaments operated with anode potentials comparable with 150 volts, and required to supply a space-charge limited current of approximately 10 milliampes per cm², the average life is given by,

$$L = 1.5 \times 10^{-5} \epsilon^{22000/T} \text{ hours.}"$$

Table XIX shows the values of the emission at $T = 1000$°K for coatings of

BaO and SrO, as calculated from the values of A and b_0 obtained by different investigators and given in TableXVI. With the exception of results by Rothe which have been discussed already, it will be observed that the emission at this temperature, varies between 10 and 100 milliamps per cm², approximately.

In commercial practice the different types of oxide coated cathodes are usually operated at an energy input of approximately 4 watts/cm² with an emission efficiency of about 50 milliamps/watt, corresponding to 200 milliamps/cm².

A survey of the data in Table XVI shows that the values of the work function for the oxides of Ba, Sr, and Ca obtained by earlier investigators are considerably higher than those observed more recently. This point has been emphasized by Espe in a special table giving the values of Φ obtained both from thermionic and calorimetric observations by various investigators since 1904. [See also Reimann and Murgoci[105] and Lowry[100]]. Espe ascribes this observed decrease in values of Φ to improved vacuum technique, since "even very slight amounts of residual gases can cause an increase in the work function." (This will be discussed in the section on "The Effect of Gases on Thermionic Emission.") Furthermore, there has naturally been a considerable improvement in the technique of preparation and activation of the cathode materials. In view of these considerations we must regard as most reliable the values deduced by Koller, Espe, Detels and Davisson. The values given by the last named investigator, since they represent carefully weighted averages of an extremely large number of emission curves, are to be regarded as probably the most accurate emission constants available at present on oxide-coated cathodes.

A determination of the work function for pure platinum, coated with oxides of barium and strontium, from emission data has been reported by C. Davisson and L. H. Germer.[107] The values obtained for Φ varied from 1.79 initially to 1.38 volts in the steady state, whereas the direct determination on the same filament by the calorimetric method (see following section) gave 1.79 volts. These observations as well as those made by all the other investigators show not only the effect of ageing on the emission constants but also that, in contrast to the results obtained with pure metal filaments, it is virtually impossible to obtain data of the same reproduceable nature. It may be, as J. A. Becker,[103] Schottky[98] and others have suggested, that the work function of a composite surface is not independent of temperature, and that therefore, "no great significance is to be put on any work functions determined in this manner," that is, from plots of log $I/T^{1/2}$ or of log I/T^2 vs. $1/T$. Nevertheless the fact that the slopes of these lines show a wide range of values is characteristic of this type of cathode.

The tacit assumption made so far in the discussion, that the base or core metal has no effect on the emission constants, is founded to some extent on the results of the early investigations of F. Deininger.[112] He observed that the work functions for the various oxides were independent of the nature of the core material. Schottky and Rothe[97] have expressed the same opinion

in their treatise (1929). This conclusion has, however, been questioned in a very recent paper by E. F. Lowry[100] as a result of an investigation on the behavior of Konel metal filaments coated with BaO and SrO. The composition of this metal has been given in a previous section and it was also pointed out in that connection that the material is a much better radiator than platinum. While no values of the thermionic emission constants are given, it is stated by Lowry that, "the same anode current may be obtained from oxide coated Konel at 775° C as can be obtained from platinum coated with the same oxides at 950° C." The theory suggested in explanation of this result will be given in connection with the discussion of activation phenomena in general.

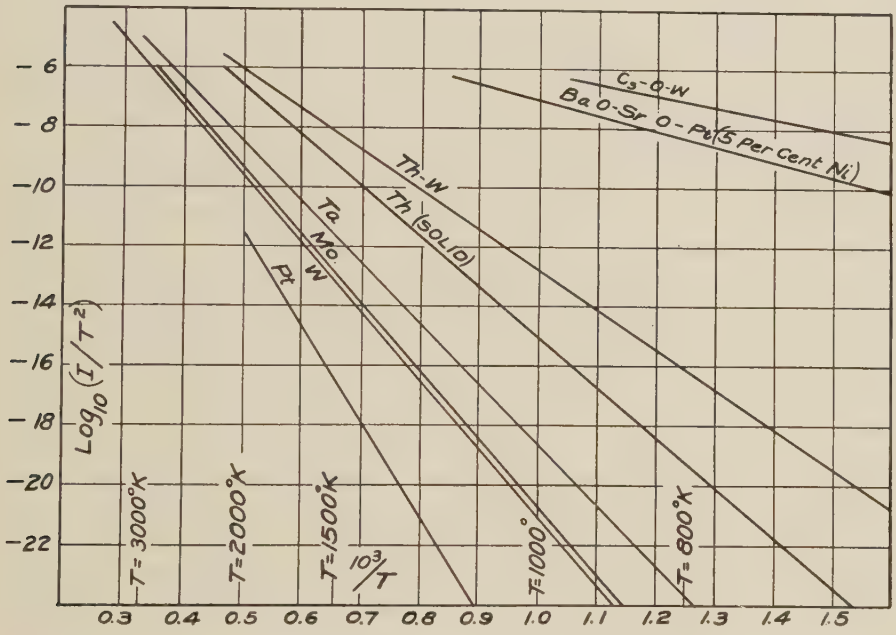


Fig. 11. Comparison of emission from coated filaments with that from pure metals.

The thermionic properties of the rare earth elements have been investigated by E. E. Schumacker and J. E. Harris.[117] Cerium, lanthanum, praesodymium, neodymium, and samarium were tested as chemically pure metals, while alloys with aluminum were tried in the case of the following metals: yttrium, europium, gadolinum, terbium, dysprosium, holmium erbium, thulium, ytterbium and lutecium. "Due to inherent difficulties in the methods that had to be used in making the measurements," the authors express the opinion that "no great accuracy can be claimed for the data" actually obtained. From the plots of log i versus $1/T$ given in the paper, it appears that the values of b_0 vary from about 40,000 to 30,000 for nearly all these elements. However, in view of the fact that the fraction of the

[117] E. E. Schumacher and J. E. Harris, J. Am. Chem. Soc. **48**, 3108 (1926).

surface active for emission probably changes in an indeterminate manner with temperature, these values may be quite inaccurate. The temperature at which an emission comparable to that of tungsten at $T=2000°$ K (1.0 milliamp/cm^2) is obtained is as low as 1200° K for thulium and is higher for the other elements. Thus the observed emission is lower in all these cases than that observed for coatings of barium and strontium oxides.

Finally, it is of interest to compare the emission from the latter with that obtained from pure metals and monatomic films on metals. Fig. 11 shows, plots of log (I/T^2) vs. $1/T$ for a few of the more important emitters ranging from Cs-O-W which has the lowest values of b (or b_0) to platinum, which apparently has the highest value of b observed up till now for a pure metal. The upper limit of each plot (lowest value of $1/T$) as shown corresponds to the melting point in the case of pure metals or to the temperature at which the surface atoms evaporate very rapidly in the case of emitting films. For example, in the case of thoriated tungsten this temperature (for maintaining $\theta=1$) is 2000° K–2100° K, and for oxide coated cathodes about 1200° K.

(D) Calorimetric determination of work function for oxide coated cathodes.

As in the case of metallic filaments, a number of investigators have attempted to determine the work function from observations on the cooling effects. The earlier investigations have been discussed by Richardson. Table XX gives a summary of the most reliable results. The third column gives the value of the absolute temperature at which Φ (cal.) was determined calorimetrically. The values under Φ (therm.) were obtained in all cases, except those of Spanner and Michel's results, from thermionic emission

TABLE XX. *Calorimetric determinations of work function for oxide coatings.*

Author	Oxides	T	Φ (cal.)	Φ (therm.)	Ref.
Schneider	CaO		3.45	3.45	(118)
W. Wilson	BaO—50% SrO —50%		1.97 2.28	2.02 2.16	(119) (120) (96)
	BaO—50% SrO—25% CaO—25%		2.39 2.54	2.34 2.59	
	CaO		3.22 3.25	3.28 3.49	
Davisson and Germer	BaO—50% SrO —50%	1064	1.79 (1.61)	1.79 (1.65)	(107)
Michel and Spanner	CaO SrO BaO BaO+SrO	1493 1458 950 1064	2.40 2.15 1.85 1.61	2.45 2.27 1.69 1.79	(121)

[118] H. Schneider, Ann. d. Physik **37**, 569 (1912).
[119] W. Wilson, Proc. Nat. Acad. Sci. **3**, 426 (1917).
[120] W. Wilson, Phys. Rev. **10**, 79 (1917).
[121] G. Michel and H. J. Spanner, Zeits. f. Physik **35**, 395 (1925).

measurements on the same filaments. Spanner and Michel's determination of Φ (cal.) are compared with Spanner's results for Φ (therm.) as given in Table XVI. Davisson and Germer calculated Φ_0 from their determinations of Φ and these are given in brackets. In view of the experimental difficulties involved in such measurements, the agreement between the values obtained for Φ by the two methods must be regarded as very satisfactory.

Michel and Spanner have suggested that the measurements on oxide coatings really yield the work function for the core metal. Schottky however, disagrees with this conclusion.

(E) Activation and emission phenomena in oxide coated filaments.

As has been mentioned previously, the phenomena in this type of filament are quite different from those observed with metallic filaments such as tungsten. The activation phenomena are similar in certain respects to those observed with thoriated and caesiated tungsten cathodes. It is therefore to be surmised that the increase in emission is due to some material diffusing to the surface during activation. The additional feature, however, of the necessity of applying plate voltage in order to obtain activation requires a further assumption. Furthermore, the large variation in emission constants obtained by different investigators and even by the same observer on filaments of the same composition and similar treatment—these observations lead to the inference that the number of factors controlling the emission must be greater than in the case of monatomic films on metals. To these characteristics of coated filaments must also be added the following characteristics:

(1) Lack of definite saturation emission.

(2) "Changes in activity with time following upon or accompanying (a) changes in temperature, (b) changes in plate potential, and (c) current sent into or drawn from the filament." (Becker).[103]

(1) *Evidence for metallic barium as "active" material.* That the active material produced by thermal treatment alone must be metallic barium has been shown by Koller[99] as follows: On deactivating a coated filament by heating at 1600° K for a short interval, the activity could be restored by glowing at 900° K. It was found that the rate of activation at the latter temperature increases with time of flashing at 1600° K, passes through a maximum and then decreases. "The explanation is that active material is formed at the high temperature but the rate of evaporation is too high to permit it to remain on the surface. At lower temperatures the surface becomes partially covered by some of the material which diffuses from inside the oxide coating. The rate at which the surface becomes covered depends on the concentration within the coating. This increases with the length of time of flashing up to the point where further flashing results in more active material being lost by evaporation than is formed." Similar observations have been made by M. S. Glass.[104]

Koller also observed that the activity is decreased by a pressure as low as 5×10^{-3} mm oxygen, and that water vapor has the same effect in "poison-

ing" the emission. Reimann and Murgoci[105] have shown that "poisoning" also occurs in presence of an electrical discharge in carbon monoxide, and Lowry finds that the activity is destroyed by overheating.

These observations are in accord, as mentioned above, with the theory that the barium forms an adsorbed layer one atom thick either on the surface of the oxide, as assumed by most investigators, or on the surface of the core material as assumed by Lowry and also by Reimann and Murgoci. It follows that the emission observed under varying conditions of treatment must be ascribed to the effect of these conditions on θ, the fraction of the surface covered with barium atoms. That the rate of evaporation of barium atoms from such a film must be considerably less than that of barium as metal follows, as pointed out by Koller, Schottky and Rothe[97] and others, from the analogy with the behavior of monatomic films of thorium and caesium on tungsten. Further evidence for this lower rate of evaporation of adsorbed barium is presented in a later section.

To these arguments in favor of barium as active material must be added the fact that in the barium azide process for the manufacture of cathodes the barium is deposited on the filament in the atomic state and that the emission observed is often actually greater than that observed at the same temperature for filaments coated with the oxides in the more usual manner. (See the comments in a previous section on the emission data obtained by Rothe).[110]

Furthermore, there are a number of observations on the emission of tungsten filaments on which metallic barium has been deposited by evaporation at relatively low temperatures of activated coated filaments. Thus from data given in Becker's paper[103] it follows that for a monatomic film of barium on tungsten the emission at $T = 1000°$ K is about 0.4 amp/cm^2, which is higher than the average emission of 0.09 amp/cm^2 at the same temperature for a regular coated filament according to Davisson's data in Table XVIII.

Reimann and Murgoci state that the emission obtained in their laboratories by Ryde and Harris from monatomic layers of barium on tungsten and of barium on tungsten covered with oxygen is of the same order of magnitude as that given by ordinary oxide coated filaments at the same temperature.

It should also be added that chemical tests on activated filaments, immediately after removal from the tube, have shown the presence of the alkaline earth elements in the metallic conditions.[101]

(2) *Thermal formation of barium.* Regarding the mechanism of production of barium there exist two divergent opinions. One view is that the metal is produced by thermal reaction between the oxide and core-material; the other view is that the barium is produced by electrolysis of the oxide. The first theory is apparently held by Koller and is supported by Lowry and other investigators. The second theory is favored by Rothe, Espe, Detels, Becker and Reimann and Murgoci.

Lowry's experiments on the activation of coated Konel metal filaments by purely thermal treatment certainly favors the chemical reaction theory, at least for this type of filament. There is some evidence, however, that

even in the case of platinum or tungsten filaments coated with oxides activation may be produced, although rather slowly, by thermal treatment alone.

In experiments in this laboratory with "unipotential" cathodes of nickel covered with oxides of barium and strontium, such as used for arc discharges in the rare gases and mercury[122] the cathodes are activated completely by high-frequency heating without application of voltage.

The theory of the formation of barium by thermal reaction has also received strong support from L. J. Davies[123] and A. Gehrts.[124] The former mentions that regular activated filaments have been produced by thermal treatment alone, while the latter advances considerable experimental evidence in favor of the same point of view, as for instance, the evaporation of barium during the process of activation by purely thermal treatment. Gehrts assumes that firstly the oxide is dissociated under these conditions, that secondly both barium and oxygen diffuse to the surface and that while the latter evaporates, the former remains on the surface as an adsorbed layer. Reaction of the oxide with core material is also considered as a possible source of barium, and in this respect the theory is similar to that advanced by Lowry. Gehrts also discusses in his paper the electrolytic formation theory of barium and advances reasons for disagreeing with this theory.

The theory suggested by V. C. Macnabb[125] may be regarded as an electrochemical variant of these views, inasmuch as he assumes that electrical bombardment of the cathode by ions, preferably those produced by discharge in CO_2, is necessary in order to produce the oxides "in an emissive state." According to this investigation, "The best source of the necessary gas to be used in processing is in the filament itself and can best be obtained by leaving on the filament a large percentage of uncombined oxides as the gas liberated from the filament is probably CO_2 or of some nature that is not poisonous to the filament." It is concluded furthermore, "That the combined layer of the oxides with the core comprise little or no part of the active part of the filament and only serve as a mechanical bond, and that the core material, if it has no chemical reaction with the coating, is not critical in aiding or preventing electron emission."

That the dissociation of the BaO by heat alone, in accordance with the equation

$$BaO \rightarrow Ba + \tfrac{1}{2}O_2$$

occurs at temperatures below the melting point of nickel would appear highly improbable. Calculations on this point by Dr. D. S. Villars and the writer lead to the conclusion that with the most favorable assumptions, the dissociation pressure at $T = 1000°$ K would be less than 10^{-40} atm. On the other hand, it is probable that there is a reduction of BaO by the metal of

[122] See publications by A. W. Hull, Trans. A. I. E. E. **47**, 753 (1928) and C. G. Found and J. Forney, ibid **47**, 747 (1928) on hot cathode neon lamps.

[123] This theory is given in the discussion of the paper by Hodgson, Harley and Pratt, reference 106.

[124] A. Gehrts, Zs. f. Techn. Physik **11**, 246 (1930).

[125] V. C. Macnabb, J. Opt. Soc. Am. and Rev. Sci. Inst. **19**, 33 (1929).

the core, which would be analogous to the reduction of ThO_2 in tungsten filaments by the tungsten atoms. It would not be necessary to have the reaction occur to any greater extent than the reduction of a very small fraction of the barium oxide present.

E. R. Wagner[101] agrees with Macnabb in assuming that activation occurs by bombardment with ions produced in CO_2. From consideration of the magnitudes of the free energies involved in the different reactions he concludes that at 1000° C the dissociation of the carbonates to oxides is practically complete, but that the reduction of BaO by Ni would not occcur to a sufficient extent to account for the observed emission.

According to the views sponsored by Macnabb and Wagner, the activation by plate voltage (regarded as proof of the electrolytic formation theory) is therefore to be regarded as essentially a removal of adsorbed oxygen by sputtering with the positive ions of the gas. Dr. A. W. Hull has also suggested to the writer similar views on this subject. However, as mentioned in the following section, it is probable that at least in certain cases formation of barium by electrolysis also occurs to a considerable extent, and that while this process does not necessarily continue during the life of the filament, it furnishes the mechanism by which the barium is formed during the activation stage.

(3) *Formation of barium by electrolysis.* As shown by Reimann and Murgoci[105] the electrical conductivity of the oxide coating increases during activation at the same rate as the thermionic emissivity. Furthermore, "both the thermionic emission and the conductivity are 'poisoned' similarly by exposure of the oxide to (a) oxygen, (b) a discharge in carbon monoxide, and (c) a discharge in hydrogen. Complete recovery of the formed condition by re-forming is possibly only a few times in succession after poisonings by (a) and (b), but any number of times after poisoning by (c)."

Horton[126] and Spanner[115] observed that the electrical conductivity (c) of the oxides of the alkaline earth metals increase with the temperature according to an exponential law of the form

$$c = \alpha\epsilon^{-\beta/T} \tag{26}$$

where α and β are constant. Reimann and Murgoci found that for activated filaments the ratio $\beta/b = 1$ approximately, where b is the emission constant. It is evident that β corresponds to a heat of dissociation into ions and Span-

TABLE XXI. *Ionizing potentials (V_i); energy of dissociation (V_d) and work functions for oxides (Spanner).*

Oxide	V_i	Φ	V_d
BeO		3.45 volts	1.54 volts
MgO	7.61 volts	3.01	1.6
CaO	6.09	2.4	1.45
SrO	5.67	2.15	1.32
BaO	5.19	1.85	1.11

[126] F. Horton, Phil. Mag. **11**, 505 (1906).

ner finds for this energy the values (expressed in electron volts) shown in Table XXI under V_d. The second column gives for comparison the ionizing potentials (V_i) of the metallic atoms, and the third column gives Spanner's values of Φ observed on the same cathodes that were used for determining V_d.

During electrolysis the barium ions (Ba^{++}) migrate to the core and after neutralization diffuse back towards the surface. Reimann and Murgoci assume that the oxygen ions (O^{--}) do not migrate although it is difficult to understand how the relatively heavier Ba^{++} ions can diffuse through the BaO lattice. Becker[103] concludes that in the initial stages an appreciable part of the current is carried by ions, but after the filament has been activated

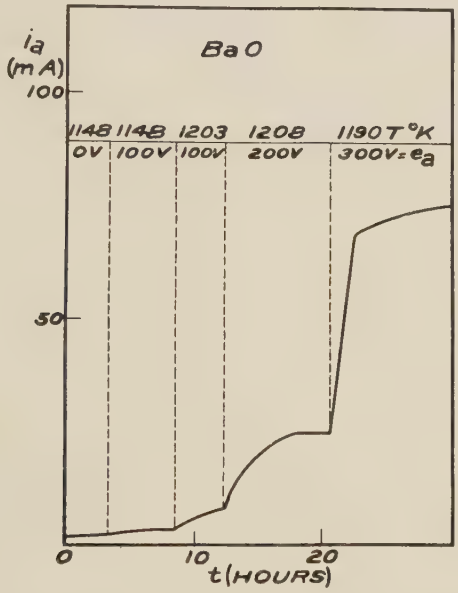


Fig. 12. Progress of activation of a BaO coating on platinum-iridium.

and barium is dispersed through the oxide (by diffusion) only a relatively small part (about 0.5 percent) of the current is carried by ions, and the rest is presumably carried by electrons through the coating.

By sending electrons into a barium oxide coated filament, Becker has shown that "the amount of barium brought to the surface by electrolysis is directly proportional to the quantity of electricity sent into the filament."

(4) *Evolution of oxygen from oxide coated cathodes.* The evolution of oxygen during activation by application of plate current was observed by Horton[126] who actually found that in the initial stages the rate of evolution of gas is within eight percent of that required by Faraday's law. Rothe[110] also observed that gas is evolved during activation by application of plate current, and the same phenomenon has been remarked by Espe[102] and Detels[108]. The latter identified the gas as oxygen by collecting it and examining the

spectrum in a discharge tube. H. A. Barton[127] has demonstrated by positive ray analysis that doubly charged oxygen molecules (O_2^{--}) are evolved during activation.

It is therefore well established that in the process of activation by application of plate voltage and drawing electrons the oxide is decomposed like an electrolyte, with liberation of oxygen at the outer surface (towards the anode) and deposition of barium presumably on the core metal. The activity of the surface is thus increased until an equilibrium state is established between evaporation of barium, diffusion of the metal from the core, and recombination of barium and oxygen on the surface.

The effect of both increased cathode temperature and increased plate voltage during activation is illustrated by Fig. 12 taken from Espe's paper which shows the activation of a BaO coating on platinum-iridium. Fig. 13 shows the manner in which the emission increased with time when a plate

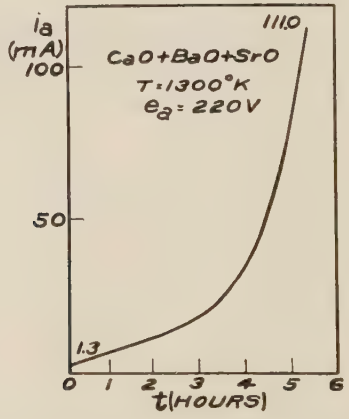


Fig. 13. Increase in emission of coated filament with 220 volts on plate.

voltage of 220 V was applied to a filament coated with a mixture of the oxides. The final emission current thus depends upon the temperature and plate voltage. A change in either of these variables is therefore followed by time changes in the emission until a new equilibrium state has been reestablished (fatigue effects).

(5) *Changes in emission constants during activation.* Before discussing in greater detail these fatigue effects observed with activated filaments, it is of interest to mention the observations on the variation in the constants A and b_0 during activation.

Espe observed that while b_0 remained fairly constant, A increased considerably. Fig. 14 shows plots of log $(i/T^{1/2})$ vs. $1/T$ taken on a SrO coating at an anode potential of 200 volts. "In curve I," he states, "the cathode was but slightly electrolyzed in advance, so that comparatively few pure metal particles could be present. Consequently, the quantity A is very low to

[127] H. A. Barton, Phys. Rev. **26,** 360 (1925).

begin with, but grows very rapidly during the measurements at higher temperatures as is brought out in the next measurement (curve II). The increase continues in the subsequent measurement (curve III) until we reach stable conditions and the curve becomes capable of reproduction. If the anode potential is raised to 300 volts, the intermediate period of adjustment is followed by a final state which gives curve IV. The latter curve can

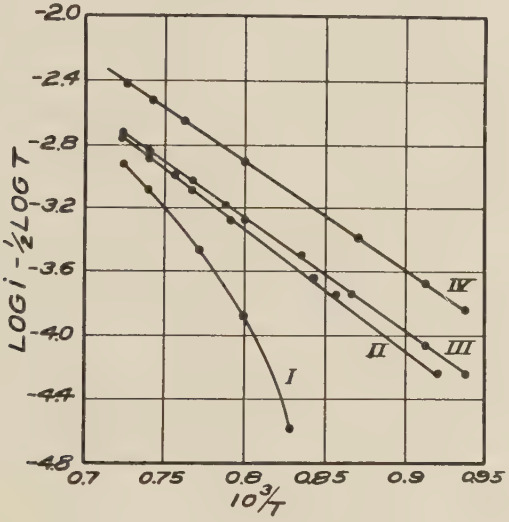


Fig. 14. Emission characteristics of a filament coated with SrO.

be reproduced and is parallel to curve III. While the curves I and II would give entirely wrong results, the curves III and IV give fully reliable data."

On the other hand, Detels has observed that both A and b_0 decrease during activation. Fig. 15 shows plots of log (I/T^2) vs. $1/T$ taken at intervals of two hours during formation with a plate current of 70 to 100 milliamperes

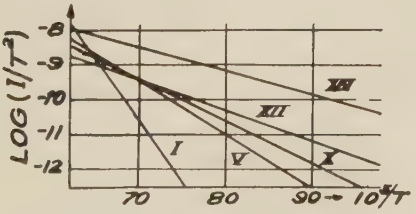


Fig. 15. Showing that both A and b_0 decrease during activation.

and voltage of approximately 100. The numbers on the plots indicate the order. Thus curve V was taken after four and a half hours' run, and the other curves at subsequent periods (up to 3 days). A plot of Φ_0 vs. log A, such as shown in Fig. 16 indicates that the decrease in log A was approximately proportional to that in Φ_0. Thus A varied from 10^{15} to 10^{16} at the beginning to 9.3×10^{-4} at the end, while Φ_0 decreased from 7 to 1.36 volts.

Similar observations were made with oxidized nickel wires covered with alkaline earth oxides. The values of A and Φ_0 during formation are given in Table XXII.

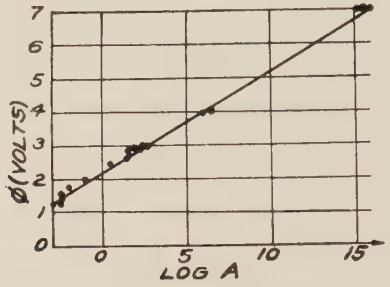


Fig. 16. Showing that during activation the decrease in log A is practically proportional to that in ϕ_0.

We shall now proceed to discuss the different processes which apparently occur in activated filaments and govern the observed electronic emissivity.

TABLE XXII. *Variation in emission constants during activation of oxide coated nickel (Detels).*

Time (hours)	Φ_0 (volts)	A (amps/cm² deg.²)
$\frac{1}{6}$	2.37	27.5
$1\frac{1}{2}$	1.83	1.24
2	1.36	0.0373
$2\frac{1}{4}$	1.36	0.0297

(6) *Evaporation of barium.* Becker[103] has shown that a pure tungsten filament adjacent to an activated coated filament maintained at a temperature above 1250° K increases enormously in activity owing to evaporation of barium from the oxide coating. Fig. 1, to which reference has been made in a previous section, from Becker's paper, shows the increase in emission from the tungsten filament with increase in θ. The temperature of evaporation was 1260° K. On glowing the tungsten filament to 1700° K, the barium was removed completely.

J. M. Eglin[116] finds that the rate of evaporation of barium from a coated filament produced by the azide process is of the order of 10^{-9} of that of barium in bulk. That is, the adsorbed barium, like thorium or calcium on tungsten, has an extremely low rate of evaporation as compared with the metal itself.

At higher temperatures *BaO molecules evaporate* along with barium atoms. C. Davisson (quoted by Arnold[96]) found that the rate of evaporation of the oxide is given by the relation

$$m = 4.6 \times 10^8 T^{1/2} \epsilon^{-46000/T}$$

where m = rate in gr/cm² sec.

C. Zwikker[127a] has also reported data obtained by Reernik which give a

[127a] C. Zwikker, Physica **8**, 241 (1928).

considerably higher rate of evaporation at the same temperature. (A plot of the two sets of data is given by Becker.)

C. Davisson and H. A. Pidgeon[128] found by evaporating barium oxide from a coated filament at 1373° K to an adjacent tungsten filament that the emission from the latter was increased by a factor 10^9 and "that the maximum emission occurs when there are approximately 2.4×10^{14} molecules BaO per cm^2 of secondary. This is reckoned to be not more than 30 percent of the number required to form a layer one molecule deep."

(7) *Evaporation and diffusion of oxygen.* According to Becker, "When current is drawn from a coated filament at a sufficiently high temperature and if this current is not limited by space charge, oxygen evaporates." The evolution of this gas was measured by the decrease in activity of an adjacent tungsten filament. In this manner it was found possible to determine the factors upon which the rate of evaporation of oxygen depends. "Such experiments," Becker states, "have shown that the rate of oxygen evaporation increases with (1) the temperature of the filament, (2) the amount of current drawn from it, (3) the plate potential, and (4) the composition or state of the oxide."

It should be mentioned in this connection that Espe and others have observed that in well-exhausted tubes with oxide coated cathodes Langmuir's space charge relation, viz., that the space current varies as $V^{3/2}$, is applicable, thus proving not only the absence of positively charged ions, but also that absence of sharp saturation for the emission is not due to the presence of residual gas. Therefore any oxygen liberated during the operation of the tube must be either negatively charged or else "cleaned up" by the barium on the surface of the cathode and on the walls of the tube (by evaporation from the cathode).

Becker has also given evidence for assuming that oxygen diffuses from the surface into the oxide and presumably recombines to some extent with the barium diffusing from the surface of the core.

(8) *Diffusion of barium in the oxide.* The barium deposited on the oxide core interface must reach the outer surface by diffusion. As Becker points out a semi-quantitative proof of diffusion is based on the life of oxide coated filaments. From the total amount of electricity carried through the core during a life of 15,000 hours he finds, on the theory that only 0.001 of the current is carried by ions, that over 9×10^6 layers of barium should be deposited at the core surface. Since there are actually only about 2×10^4 layers of oxide on a filament and only a portion of these disappear during the life, it is concluded that "by far the largest part, about 0.998, of the barium which is produced by electrolysis must diffuse into the oxide, and must recombine with oxygen to form barium oxide."

Reimann and Murgoci have suggested that the barium deposited on the core diffuses back to the surface along the surfaces of the oxide crystals, in analogy with the diffusion of thorium along grain boundaries in the case of

[128] C. Davisson and H. A. Pidgeon, Phys. Rev. **15**, 553 (1920).

thoriated tungsten filaments. This *theory of barium circulation*, that is, that the barium migrates to the core as ions and returns to the surface along the surfaces of the oxide crystals, leads to the conclusion, based on the observed life of oxide coated filaments, that the average life of an adsorbed atom of barium on the surface is of the order of 10^{-3} seconds. "This result," they point out, "is in excellent agreement with the value (also 10^{-3} seconds) calculated by W. Schottky[129] from J. B. Johnson's[130] measurements of the 'flicker' of oxide cathodes. Incidentally it is plain that Schottky's result when considered in conjunction with the great length of life of oxide cathodes, constitutes independent evidence in support of an essential part of our theory, viz., the repeated appearance of the same particles of barium at the emitting crystal surface."

On this basis, "the concentration of the film is determined, in any given case, by the requirement of statistical equilibrium between the rate of removal of the barium atoms from the emitting surface (as ions) and the rate of their replenishment by diffusion." Also this theory offers an explanation of the forming and poisoning of the conductivity under conditions in which the emission is also formed or poisoned. "Conductivity and emission are formed and poisoned together, simply because they both depend on the same thing, viz., the presence of uncombined barium at the surface of the crystals."

(9) *Effect of oxygen on the activity.* As Becker remarks, "It is a well known fact that when current is drawn from a coated filament the activity of this filament changes rapidly at first and then more and more slowly. The magnitude and even the direction of the change in activity depend greatly upon the composition of the oxide, the plate potential, the degree of break-down (activation) and the temperature of the filament. For a well broken down (activated) filament we have invariably found that the activity decreases rapidly and then more and more slowly. . . . . It was shown above that if the temperature and the plate potential are high, oxygen is formed by electrolysis and evaporates. Presumably, therefore, the decay in activity is caused by oxygen which has come to the surface and some of which stays there because the temperature does not remove it rapidly. Because of its electron affinity, some of the oxygen should be negatively charged and if the oxygen stays on top of the surface layer of barium, it should produce electrical fields which hinder the escape of electrons. Consequently the activity should decrease. If more and more oxygen is brought to the surface, the rate at which it disappears from the surface either by evaporation or diffusion should increase rapidly until the rate of disappearance is equal to the rate of production. We thus have a satisfactory picture of the cause of the decay in activity when current is drawn."

Becker does not believe that oxygen on the surface combines with the barium, since a deactivated filament (due to oxygen on the surface) can be

[129] W. Schottky, Phys. Rev. **28**, 74 (1926).
[130] J. B. Johnson, Phys. Rev. **26**, 71 (1925).

restored by glowing at a high temperature, such as 1200° K, at which temperature the oxygen leaves the surface by evaporation.

It is difficult to believe, however, in view of the fact that adsorbed oxygen on thorium, cerium and other rare earth metals forms layers which are stable even at extremely high temperatures, that films of oxygen on barium can be removed by heating at such a relatively low temperature. Is it not possible that in Becker's experiments the restored activity is due to increased rate of diffusion of barium from inside the coating?

Other observations on filaments which have not been activated lead Becker to conclude that, "From analogy with caesium on tungsten it is to be expected that a little oxygen increases the activity, while more oxygen decreases it. Whether the activity is increased or decreased depends upon just where the oxygen is located with reference to the surface barium atoms. If the oxygen is beneath the barium layer, the activity is raised because the negative oxygen induces more positive barium ions and the moment of the additional barium ions and their images more than counterbalances the moment of the oxygen ions and images. If, on the other hand, the oxygen is on top of the surface layer of barium, the activity is decreased because now the moment due to the oxygen is greater than the moment due to the barium.

"By combining the hypothesis with the experimental facts we may draw the following conclusions: oxygen is deposited by electrolysis just beneath the partial surface layer of metallic barium or strontium. If the temperature is high enough some or all of this oxygen rather rapidly arrives at the top of the barium layer where it forms an outer oxygen layer which is a more stable layer than the inner oxygen layer. If the temperature is low enough the first oxygen that is deposited stays in the inner layer. As the oxygen concentration increases, the field tending to pull electrons and negative oxygen ions toward the plane of the barium ions increases and some of the oxygen reaches the top of the barium layer. At the peak the factors tending to increase the activity are balanced by those tending to decrease it. Beyond the peak the net result of electrolysis is the increase in concentration of the outer layer of oxygen. We believe that some of the oxygen also diffuses inward into the oxide. If the filament is glowed either before or after the peak, oxygen leaves the inner layer; it reaches the outer layer or diffuses into the oxide."

(10) *Theories of transfer of electrons in coating and effect on work function.* Reimann and Murgoci conclude from their experiments that, "The coating is probably in very imperfect contact with the core metal, so that the space current passes from core metal to coating mainly in the form of thermionically emitted electrons." Thus the electrons are emitted from the barium *adsorbed on the core.* On the basis of this theory, "it is seen that it is *á priori* by no means certain whether the Richardson straight line represents the thermionic properties of the contaminated core-metal or those of the coating. Either the core metal emits electrons at a greater rate than the coating can re-emit them, in which case a fraction of the electrons emitted by the core-metal returns to it, or else the coating emits more electrons than it receives

from the core metal, so that it re-absorbs a fraction of them. In other words, the thermionic emission of the cathode as a whole is limited either by the coating or by the core metal, whichever of the two systems in series has the lesser thermionic emissivity.

"It appears that, at least in certain cases, it is the contaminated core-metal, and not the coating, by which the emission is limited. Since, according to our assumption, the coating receives as many electrons from the space that separates it from the core-metal as it permanently loses, by re-emission, from its outer surface, there can be no net heat loss from the coating, and the only cooling that the cathode will suffer by its electron emission will be that due to the evaporation of electrons from the contaminated core metal. The work-function of the cathode, calculated from the latent heat of evaporation of electrons, will therefore in any case be that of the contaminated core-metal. Occasional comparisons have been made, e.g., by Davisson and Germer,[107] between the work-function measured in this way and that determined from the slope of the Richardson straight line, and the two determinations have been found to be in good agreement. But we have already seen that the slope of the Richardson straight line corresponds to the work-function of whichever of the two systems (core metal or coating) has the lesser emissivity. Thus in those cases where there is agreement between the work-functions determined by the two methods, we may conclude that, if the coating is indeed in poor contact with the core metal, it is the core metal which limits the emission. This conclusion accords well with the fact of experience that the nature of the core metal has a considerable influence on the thermionic properties of coated cathodes."

On the other hand, if the unit of barium circulation in the coating is made large enough, the emission becomes limited by the coating. The gradual increase in size of the crystals of oxide during life would thus lead to a decrease in emission because of reduction in the number of grain boundaries along which diffusion of barium may occur. Under these conditions, "the temperature variation of the emission may not be a measure of the true work function at all, the surface concentration of barium being different at each temperature. The relative importance of this falsification of the work function measurements will depend upon the size of the units of barium circulation."

Lowry's views on the mechanism of electron transfer also involve the hypothesis that the *barium forms an adsorbed layer on the core.* Therefore, the electrons emitted from the latter must diffuse through the interstices in the oxide coating, "which, however, does offer some impedance to their progress." As Lowry remarks, "These coatings are very porous. When viewed through a microscope they have the appearance of a light fluffy fall of snow. . . . . Suppose now that this coating should become more compact so that its pores are partially closed. The natural consequence would be that the apparent activity of the filament would be measurably impaired. This is precisely the type of phenomenon that does occur. It is well known that *glowing the filament for a comparatively short interval of time at a con-*

siderable overvoltage will practically ruin its emission. It has been the author's experience, at least, that a filament which has been so mistreated cannot, in general, be reactivated to anything like its original characteristics. It is also true that such filaments will, after such reactivation as is possible, *yield large emissions if very high anode voltages are applied.* The net result of such an overheating of a filament is essentially to increase the impedance of the tube by an appreciable amount. This we believe to be caused by a partial closing of the pores of the coating. Preliminary microscopic examination of some such filaments actually showed a sintering or glazing of the coating until it had the appearance of a crackled enamel.

"Exactly similar behavior is shown by filaments whose activity has gradually decreased during a long period of use. Such worn out filaments always retain an appreciable amount of coating. In fact, a fairly close inspection shows that they still possess nearly as thick a layer of oxide as they had initially. In this case also large emissions may be drawn if sufficiently high anode voltages are applied.

"If the loss of activity in these cases were due solely to a loss of metallic barium due to vaporization, it should be possible to cause these filaments to regain all their initial activity by electrolysis or by positive ion bombardment. As stated above, it is rarely, if ever, that it is possible to do this.

"*The process of electron diffusion through a porous*, poorly conductive coating may also explain the *phenomenon of non saturation* which is commonly met with in this type of cathode. It follows logically that large numbers of electrons will become enmeshed in the pores of the coating or occluded on the surface of the particles of poorly conducting material composing it. It is quite possible also that the numbers of electrons so entrapped will vary somewhat with the anode voltage applied. This process will give rise to an effect similar to a negative space charge within the confines of the coating itself. The space currents obtainable then at any given anode voltage and filament temperature represent the number of electrons per second it is possible to drag out of or through this negative charge built up within the coating layer. The result will be that much higher anode voltages will be required to approach a saturation than if this pseudo space charge did not exist. At temperatures in the neighborhood of 1000° C the electron emission from these filaments is so enormous that it is next to impossible to obtain anything approaching saturation owing to the very considerable heating effects of the emission itself. At these temperatures also the conductivity of the coatings will be increased appreciably, affording a better opportunity for the electron charge to leak off."*

Thus Lowry's arguments for assuming that barium is adsorbed on the core rather than on the oxide are two fold: (1) the fact that the nature of the core affects the emission, since at the same temperature the emission from coated Konel filaments is considerably greater than that obtained from coated platinum filaments, (2) the observed decay in emission during life

* The italicized portions are due to the writer.

can be interpreted only as due to a gradual sintering of the coating which results in increased "impedance" to the passage of electrons from the barium film on the core.

Concluding remarks on oxide coated cathodes

It is evident from the foregoing remarks that there are a great many contradictory opinions regarding the interpretation of activation and emission phenomena observed with this type of cathode. That the active source of electrons is a monatomic film of barium is apparently certain, but one school holds the opinion that the barium is adsorbed on the surface of the oxide, while the other school advances equally cogent reasons for believing that the adsorbed layer is formed on the core underneath the oxide.

Regarding the process of activation, there are three different suggestions (1) electrolysis of the oxide, (2) thermal reaction of the oxide with core material, and (3) bombardment by positive ions. It would seem most probable that the exact mechanism of activation actually varies with the composition of the core material and the manner in which the oxide is deposited on this core, so that while certain types of filaments require anode voltage for activation, with or without the presence of positive ions, other types can certainly be activated by purely thermal methods.

The similarity, at least in certain respects, between the phenomena observed with oxide coated filaments and those observed with oxidized tungsten filaments on which caesium has been deposited, on the one hand, and with thoriated tungsten on the other, is certainly significant. An analogy to oxide-coated cathodes may also be found in the phenomena observed with certain types of photoelectric surfaces, such as caesium on oxidized silver, or copper oxide on copper. In a very interesting paper, Wo. Ostwald[131] has pointed out that combinations of metal+oxide, or metal+alkali or alkaline earth oxide act not only as sources of electrons photoelectrically, but also as rectifiers and as contact catalysts. He regards all these as "disperse systems possessing low work functions for electron emission" which is presumably due to the strong field of force exerted by the oxygen atoms. These ideas are certainly extremely suggestive, and should promise an incentive for further investigation in this field.

VII Effect of Gases and Positive Ion Bombardment on Electron Emission

The earlier investigations (1903–1913)[132] dealt mainly with the effect of a hydrogen atmosphere on the emission from platinum. Since decrease in pressure and degassing of the metal led to a decrease in emission, it was concluded that the emission phenomenon itself is due to chemical reactions in presence of residual gases. Langmuir's investigations (1913 and subsequently) showed that in case of tungsten, improvement in vacuum actually increased the emission to a perfectly definite value, if care was taken to

[131] Wo. Ostwald, Kolloid Zeits. **51**, 370 (1930).

[132] These are discussed very fully by Richardson, loc. cit., Chap. 4.

eliminate at the same time effects due to space charge. Langmuir also studied the effect of low pressures of various gases on the emission from tungsten, and subsequent investigators have reported results observed with other emittors.[133] While a number of these observations have been mentioned incidentally in previous sections in connection with the discussion of emission constants, they are discussed more fully in the present section.

The effects of gases on the emission are of a two-fold nature. Firstly, they may form *adsorbed monomolecular* or *monatomic* films on the surface of the cathode. Secondly, and especially with anode voltages in excess of a certain critical value, (which varies with the nature of the gas) positive ions are formed which sputter atoms off the surface of the emitter and may thus affect the emission considerably.

The theory of the formation of monomolecular films was proposed by Langmuir in 1914[134] and was elaborated more completely in a subsequent paper, published in 1918,[135] from which the following remarks are of interest: "When gas molecules impinge against any solid or liquid surface, they do not, in general, rebound elastically, but condense on the surface, being held by the field of force of the surface atoms. These molecules may subsequently evaporate from the surface.[136] The length of time that elapses between the condensation of a molecule and its subsequent evaporation depends on the intensity of the surface forces. Adsorption is the direct result of this time lag. If the surface forces are relatively intense, evaporation will take place at only a negligible rate, so that the surface of the solid becomes completely covered with a layer of molecules. In cases of true adsorption this layer will usually be not more than one molecule deep, for as soon as the surface becomes covered by a single layer, the surface forces are chemically saturated. Where, on the other hand, the surface forces are weak, the evaporation may occur so soon after condensation that only a small fraction of the surface becomes covered by a single layer of adsorbed molecules."

The fraction of the surface covered, denoted by θ, is therefore a function of the rates of formation and of destruction of the film. It varies with the temperature, nature of both gas and solid surface, and with the rate of positive ion bombardment. Hence the amount by which the electron emission is affected depends upon the value of θ for the particular adsorbed layer. The phenomena thus resemble those described already in connection with the discussion of monatomic films of thorium, caesium and barium.

The effect of gases on the emission from *tungsten* has been discussed very fully by Langmuir.[137] Table XXIII illustrates the results obtained for different gases. As has been mentioned previously, oxygen decreases the

[133] These have been summarized by Schottky and Rothe in their treatise, pp. 204–210.

[134] I. Langmuir, Phys. Zeits. **15,** 520 (1914).

[135] I. Langmuir, J. Am. Chem. Soc. **40,** 1361 (1918).

[136] This condensation-evaporation theory is more fully discussed in Phys. Rev. **8,** 149 (1916).

[137] To the references given above should be added the review in J. Am. Chem. Soc. **38,** 2221 (1916).

TABLE XXIII. *Effect of gases on emission from tungsten.*

Gas	Pressure in mm Hg	a	$b\times10^{-4}$	References
Vacuum	0.00007	1.62×10^{-2}	5.55	(8)
H_2	.012	2.58×10^{2}	8.25	(8)
	.007	3.63×10^{9}	11.5	
	.0017	3.67×10^{7}	10.5	
	.0005	2.05×10^{3}	8.5	
N_2	.002	10.5	7.32	(8)
		7.95×10^{-1}	6.82	
O_2		3.25×10^{4}	9.43	(8)

	A	$b_0\times10^{-4}$	
O_2	5×10^{11}	10.7	(10*a*)

emission from tungsten enormously ($1:10^{-5}$). It forms adsorbed layers which are quite stable up to temperatures of 1800° K.

In the case of hydrogen, the observed effects are probably due to the indirect formation of oxygen. Thus the hydrogen reduces traces of oxides on the leads and this is dissociated in presence of the heated filament to form oxygen which is adsorbed on the surface. This reaction has long been familiar in lamp manufacture as a cause of early blackening.

"Nitrogen has no effect on the electron emission from tungsten if low voltages are used in making the measurements of the thermionic current. But if voltages high enough to cause ionization of the nitrogen are employed, then the tungsten is attacked by rapidly moving positive nitrogen ions, and simultaneously, the electron emission is greatly decreased, indicating that a film of adsorbed nitrogen is formed. If the filament is now allowed to cool to room temperature and the remaining nitrogen is pumped out, and the filament is then heated again to the same temperature as before, it is found that the electron emission is at first the same as it was in the presence of nitrogen, showing that the electron emission is determined not by the gas around the filament, but by the film adsorbed on it. Gradually the adsorbed film distills off, and the electron emission returns to its normal value."[138]

Carbon monoxide and carbon dioxide have the same effect as oxygen. Hydrocarbon vapors give up carbon to heated tungsten filaments with formation of carbides.[139] As a result the emission is decreased to approximately one third. As mentioned by Langmuir, "A filament which has taken up carbon is no longer sensitive to the effect of bombardment by positive nitrogen ions. In other words, nitrogen has no effect on the electron emission from such a filament even if high anode voltages are used. Still more remarkable is the fact that oxygen instead of lowering the electron emission of a carbon-containing filament gradually raises it. The carbon is then gradually oxidized and the electron emission returns to the value corresponding to

[138] However, some very recent experiments (which will be described in a forthcoming publication) have led Dr. Langmuir to the conclusion that the adsorbed film is formed, not by nitrogen, but by oxygen liberated from oxides on the filament leads by active nitrogen.

[139] M. R. Andrews, Jr. Phys. Chem. **27**, 270 (1923); M. R. Andrews and S. Dushman, Jr. Phys. Chem. **29**, 462 (1925).

pure tungsten. Then when the carbon has all been removed, the oxygen begins to attack the tungsten and the electron emission suddenly decreases to one percent or less of its former value. After the excess oxygen has been consumed it is found that the electron emission is again sensitive to the effect of bombardment by positive nitrogen ions, the thermionic current decreasing as the voltage is raised."

The effect of oxygen on activated *thoriated tungsten* is even more permanent than in the case of pure tungsten. Only by flashing to 2900° K can the thorium oxide be volatilized. It is then necessary to treat the filament at 2300° K in order to reactivate it. The effect of oxygen on a monatomic film of thorium occurs at such low pressures that it was only in presence of the extremely high vacuum produced by volatilizing tungsten with the bulb immersed in liquid air that Langmuir was able to obtain thorium emission in his first experiments. Subsequently it was found that the residual oxygen could be eliminated by other "getters" such as potassium, magnesium, calcium, etc., volatilized on the bulb walls, and this method made it possible to produce commercial radio tubes using this type of cathode.

It is instructive to realize that even at the lowest pressures ordinarily attained under best exhaust conditions the rate at which oxygen can condense on any surface is extremely high. According to the kinetic theory of gases, the number of molecules striking 1 square centimeter per second is given by

$$n = \frac{1}{4} n_0 \left(\frac{8RT}{\pi M}\right)^{1/2}$$

where n_0 = number per cm³ at temperature T,

M = molecular weight of gas.

In the case of oxygen at 25°C and 1 mm pressure, $n = 3.623 \times 10^{20}$ molecules/cm² sec. Assuming, as deduced by Langmuir, that for thorium films on tungsten, the number of atoms per cm², $N_0 = 7.56 \times 10^{14}$, and that 1 molecule of oxygen is adsorbed for every atom of thorium, it is evident that even at 2×10^{-6} mm pressure, *it would require only one second* to cover a thorium film completely with oxygen. Hence in order to obtain a life of several thousand hours, as is actually obtained with such filaments, the residual oxygen pressure must be less than 10^{-9} dyne/cm² or 10^{-15} atmospheres.

It is this type of phenomenon that makes it so difficult to obtain emission data from absolutely clean surfaces of metals. Even if the pressure is reduced to an extremely low value, and the surface has been cleaned by heating at a high temperature, there is no assurance that on cooling to lower temperatures no adsorbed layers will be formed by condensation of the residual gases.[140]

In the case of *thoriated tungsten*, Kingdon[10] has shown that nitrogen at a pressure of 0.0001 to 0.005 mm causes an approximately five-fold increase is emission.

The effects of oxygen and other gases on *molybdenum* are very similar

[140] S. Dushman, "High Vacuum."

to those observed with tungsten. Since MoO_3 is less volatile than WO_3, it is much more difficult to obtain clean molybdenum surfaces than in the case of tungsten. In this connection it should be observed that according to the calorimetric determinations made by Van Voorhis[141] the work function for molybdenum is apparently between 0.75 and 1.00 volt higher in presence of nitrogen and argon.

Tantalum oxide also is extremely non-volatile and therefore the emission from tantalum wire which has not been heated to a high temperature is very low. R. Suhrmann[142] observed that continued heating of a tantalum filament led to a decrease in emission. The emission constants obtained for filaments in different states of degassing are shown in Table XXIV.

TABLE XXIV. *Effect of gas on emission from tantalum.*

	a	$b\times10^{-4}$	Φ (volt)
Partly degassed	1.19×10^{5}	3.64	3.14
More completely degassed	3.61×10^{4}	3.64	3.14

According to H. Simon[143] hydrogen increases the emission from tantalum, which would account for Suhrmann's observations. Schottky[144] points out that tantalum becomes brittle in hydrogen, and that its resistance changes. Apparently a hydride is formed on the surface and Φ is thereby decreased, as in the case of platinum.

The increase in emission from *platinum* in presence of hydrogen has been mentioned above. The earliest observations were made by H. A. Wilson and O. W. Richardson, while more recent investigations have been carried out by L. Lockrow,[145] R. Suhrmann,[146] and L. A. DuBridge.[147] The last two have also determined the effect of gas on the photoelectric threshold frequency (ν_0). The concurrent values of emission constants given in Table V, especially the observations of DuBridge, show that the emission for degassed platinum is extremely small as compared with that for platinum in hydrogen. Table XXV gives additional values observed by Suhrmann and Lockrow for filaments which were only partly outgassed.

TABLE XXV. *Effect of gas on emission from platinum.*

Author	a	$b\times10^{-4}$	A	$b_0\times10^{-4}$	Φ_0
Suhrmann	7.59×10^{8}	5.79	(less degassed)		4.99
	1.21×10^{7}	5.31	(more degassed)		4.57
Lockrow	3.02×10^{3}	4.60	1.36×10^{-2}	4.40	3.87
	9.90×10^{3}	5.20	4.46×10^{-2}	5.00	4.31
	9.7×10^{8}	5.84	(in hydrocarbon vapor)		

[141] Van Voorhis, Phys. Rev. **30**, 318 (1927).
[142] R. Suhrmann, Zeits. f. Physik **13**, 17 (1923).
[143] H. Simon, Zeits. f. techn. Phys. **8**, 434 (1927).
[144] Schottky, Handbuch, p. 209.
[145] L. Lockrow, Phys. Rev. **19**, 97 (1922).
[146] R. Suhrmann, Zeits. f. Physik **13**, 17 (1923).
[147] L. A. DuBridge, Phys. Rev. **29**, 451 (1927); **31**, 236 (1928).

Suhrmann[148] observed that with increase in period of heat treatment, ν_0 increased from a minimum value corresponding to $\Phi_0 = 4.14$ to a constant maximum value which corresponded to $\Phi_0 = 4.62$. In the case of silver, a similar decrease was noted from $\Phi_0 = 4.72$ to $\Phi_0 = 4.28$, while in that of gold the threshold frequency showed only slight variation from $\Phi_0 = 4.72$. The observations made by DuBridge on the effect of gases on the photoelectric threshold frequency for platinum are much more complete and are discussed in a subsequent section.

The progressive decrease with ageing in the constants A and b_0 for *zirconium* and *hafnium* observed by C. Zwikker[149] have been mentioned previously. Traces of oxides cause an increase in both constants as compared with those for pure metals. The same conclusion holds valid for the alkali and alkaline earth metals.

The effect of oxygen, carbon monoxide and dioxide in "poisoning" the emission from fully activated *oxide coated cathodes* has been discussed already. It is evident that any gas which tends to form an adsorbed layer on top of the barium atoms must tend to decrease the emission. On the other hand, if this gas forms a film *under* the barium, it may increase the emission. This probably accounts for so many discrepant observations recorded by different investigators on the effect of oxygen on the emission from this type of cathode. No systematic investigations have been published on the effect of gases on the emission constants from coated filaments, but probably the variation in emission actually observed is to be ascribed, partly at least, to effect of residual oxygen in the bulbs. Koller[99] has shown that in presence of even 0.005 mm oxygen, A is increased from 1.07×10^{-3} to 4.27×10^{3}, and b_0 from 12,100 to 35,900.

TABLE XXVI. *Sputtering data for thorium on tungsten in various gases.*

Ion	V_0	n at 150 volts
H	>600	0.000
He	(about) 35	.00015
Ne	45	.023
Ar	47	.080
Cs	52	.084
Hg	55	.044

The effect of bombardment by *positive ions* of nitrogen on the emission from pure tungsten has been mentioned already. On the other hand, mercury, argon and other rare gases have no effect whatever on the emission. In fact, positive ions of these gases sputter off any films due to adsorbed films of oxygen or other gases, so that in presence of traces of mercury, oxygen has no appreciable effect on the emission.

The removal of thorium from monatomic films on tungsten, by positive ion bombardment, has been investigated by K. H. Kingdon and I. Lang-

[148] R. Suhrmann, Zeits. f. Physik **33**, 63 (1925).
[149] C. Zwikker, Phys. Zeits. **30**, 578 (1929).

muir.[105] While hydrogen ions produce no sputtering even up to 600 volts energy, ions of A, Cs, Hg and Ne all start to sputter the thorium ions at about 50 volts energy. Table XXVI shows the results obtained with ions of 150 volts energy. V_0 denotes the threshold voltage for sputtering, and n denotes the ratio of the number of thorium atoms removed to number of ions. In these experiments the thoriated tungsten filament was maintained cold.

A. W. Hull[150] has observed that with hot filaments and large positive ion currents, the threshold voltages are somewhat lower, viz: 27 volts for Ne, 25 for A and 22 for Hg. "These values are independent of temperature over the range 1900–2300°K." Fig. 17, taken from Hull's paper, shows the char-

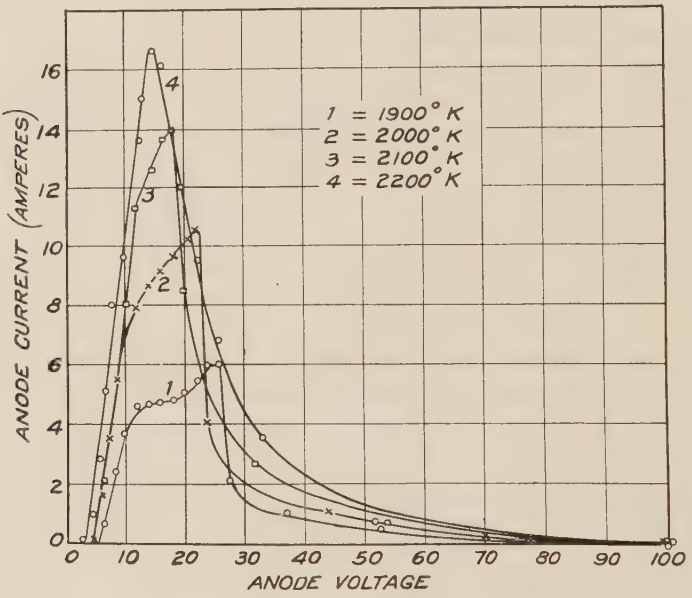


Fig. 17. Characteristics of a thoriated filament in mercury vapor at 0.005 mm pressure.

acteristics of a thoriated tungsten filament in mercury vapor at about 0.005 mm. At low voltages the electron current is limited by space charge. As the voltage is increased above the ionizing potential of the gas, the emission increases rapidly at first, but decreases as soon as the voltage rises above the threshold value for sputtering.

Similar observations have been made by Dr. Hull and other investigators in this laboratory, with oxide coated cathodes of the "unipotential" type. In all cases "disintegration (of the emitting film) does not occur when the voltage is below a critical value," and it is because of the existence of this phenomenon that it has been found possible to construct gaseous discharge tubes using hot cathodes, with a consequent extremely low cathode drop.[151]

[150] A. W. Hull, Trans. A. I. E. E. **47**, 753 (1928).

[151] Hull, Reference 150, C. G. Found and J. Forney, Trans. A. I. E. E., **47**, 747 (1928).

VIII. Photoelectric Threshold and Thermionic Work Function[152]

The maximum kinetic energy, $(\frac{1}{2}mv^2)$, of electrons emitted from a surface by monochromatic radiation of frequency ν, is given by Einstein's relation

$$\tfrac{1}{2}mv^2 = h\nu - P = h(\nu - \nu_0) \tag{27a}$$

where $P = h\nu_0$ is the work which must be done by the electrons in overcoming surface forces. The frequency, $\nu_0 = P/h$, is known as the *threshold frequency*, and is related to the work-function, Φ_0, by the equation

$$\Phi_0\epsilon = h\nu_0$$

that is

$$\Phi_0 = \frac{\nu_0}{8106 \times 3 \times 10^{10}} = \frac{12336}{\lambda_0} \tag{28}$$

where Φ_0 is expressed in volts, and λ_0 is the wave-length in Angstroms.

The energy of the emitted electrons is determined by measurements on the potential, V, required to retard them to zero, and by plotting V against corresponding values of ν, the value of ν_0 is determined from the intercept on the axis of ν, in accordance with the relations

$$\begin{aligned} V\epsilon &= \tfrac{1}{2}mv^2 \\ &= h(\nu - \nu_0) \end{aligned} \tag{27b}$$

While there have been published a large number of results for threshold frequencies for different metals, most of these results were obtained on sur-

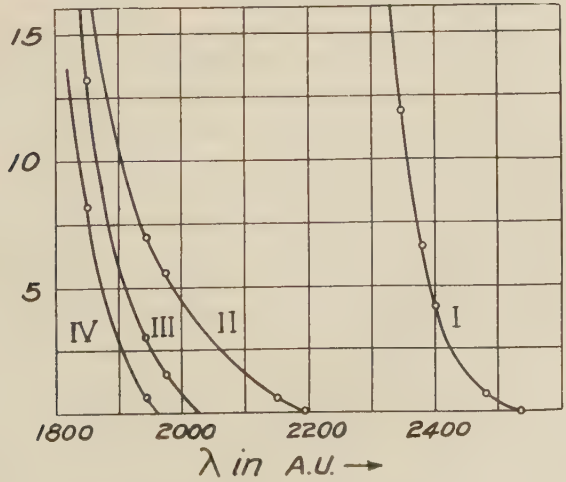


Fig. 18. Photoelectric emission of platinum as function of wave-length for various stages in outgassing.

[152] The writer wishes to acknowledge his indebtedness in this connection to Dr. A. L. Hughes and Dr. L. A. DuBridge for permission to use the material in the chapter on this subject in their forthcoming volume on "Photo-electricity."

faces which were not cleaned and only in a relatively few cases have attempts been make to measure the photoelectric frequency, ν_0, and the thermionic work function, Φ_0, on the *same* surface.

By far the most careful investigation of this nature has been carried out by L. A. DuBridge[153] on platinum. Figs. 18 and 19 show the photoelectric and thermionic curves respectively for the metal during successive stages of outgassing by heating. Curve I, taken with a fresh specimen shows a thres-

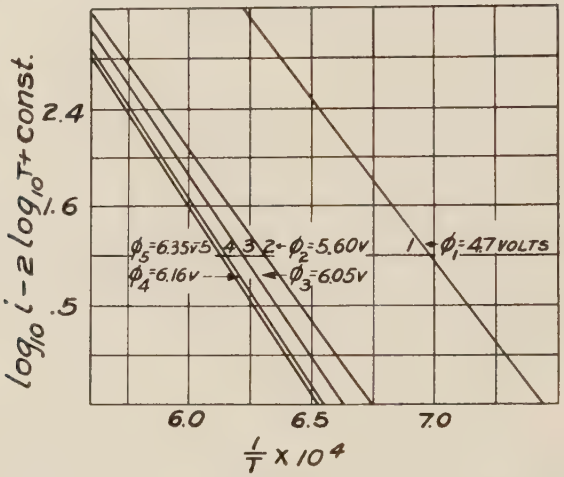


Fig. 19. Thermionic emission of platinum during various stages in outgassing.

hold at 2540A (corresponding to 4.86 volts). The value of Φ_0 under similar conditions was 4.7 volts. Continued degassing caused a decrease in λ_0 and increase in Φ_0, and the final values as shown in curves *V* and 5 are 1962A (corresponding to 6.30 volts) and $\Phi_0 = 6.35$ volts. This result thus established the conclusion that for a *thoroughly degassed metal the photoelectric and thermionic work-function are identical.*

These measurements, together with those obtained by other investigators are given in Table XXVII. The numbers in brackets give the corresponding references. In the following paragraphs these results are discussed at further length. The values observed by M. J. Martin[154] on molybdenum are low. The only explanation that can be suggested is that some impurity having a lower value of Φ_0 was present in the metal and was not eliminated at the highest temperatures to which the metal was heated in these experiments (approximately 1800°K, whereas in Dushman's experiments the metal was heated to about 2500°K). Martin also observed that "the photoelectric sensitivity showed a decrease with rise in temperature in the range from 0°C to 1000°C."

A. H. Warner[155] investigated the photoelectric behavior of degassed tung-

[153] L. A. DuBridge, Phys. Rev. **31,** 236 (1928); **29,** 451 (1927).
[154] M. J. Martin, Phys. Rev. **33,** 991 (1929).
[155] A. H. Warner, Phys. Rev. **33,** 815 (1929); Proc. Nat. Acad. **13,** 56 (1927).

sten. The long wave limit was found to be 2570 ± 50A, (4.80 volts), and was independent of temperature. This is to be compared with the thermionic value, 4.52 volts.

TABLE XXVII. *Photoelectric determination of* Φ_0.

Metal	Φ_0 (in volts) Photoelectric	Thermionic
Pt	6.30 (153) 4.62 (160) 4.40 (161)	6.27 (153) T*
Mo	3.22 (154) 4.33 (161)	3.48 (154) 4.44 (T)
W	4.80 (155) 4.52 (161)	4.52 (T)
Ta	3.93 (161)	4.07 (T)
Fe	4.72 (156) 4.2 (161)	
Ni	4.12 (161)	
C	4.82 (161)	
Ag	4.30 (160) 3.85 (161)	4.08
Au	4.72 (160) 4.71 (161)	4.42
Al	3.57 (161)	
Zn	3.68 (161)	
Hg	4.53 (157) (158)	
Sn—β	4.50 (159)	
—γ	4.38 (159)	
—l	4.21 (159)	

* T refers to values of ϕ_0 in Table IV.

A. B. Cardwell[156] observed the long wave limit for thoroughly outgassed iron as between 2580Å and 2652Å. This corresponds to a work function of 4.72 ± 0.07 volts. The variation in photoelectric sensitivity was found to be quite complex and is probable related to the change in crystal structure of the metal.

C. B. Kazda's[157] determination of the value of λ_0 for mercury has been checked by H. K. Dunn,[158] while A. Goetz's[159] determinations on the β-(Tetragonal), γ-(hexagonal) and liquid (l) tin were carried out under very carefully controlled conditions.

[156] A. B. Cardwell, Proc. Nat. Acad. **14,** 439 (1928).
[157] C. B. Kazda, Phys. Rev. **22,** 523 (1923).
[158] H. K. Dunn, Phys. Rev. **29,** 693 (1927).
[159] A. Goetz, Phys. Rev. **33,** 373 (1929).

R. Suhrmann[160] and S. C. Roy[161] have used a somewhat indirect method for determining ν_0. The theory of this method was first developed by Richardson,[162] and has also been discussed by W. Wilson[163] and A. Becker.[164] The photo current (I_p) is determined as a function of the temperature of the source of illumination and log (I_p/T^2) is plotted as a function of $1/T$. The slope of the straight line gives the value of Φ_0 corresponding to ν_0.

R. Suhrmann determined Φ_0 for platinum, silver and gold. His value for the first metal (4.62 volts) is much lower than that of DuBridge. The values for the other two metals are given in the table and also the values obtained by Goetz from thermionic data (see section V) S. C. Roy s values for Ta, W, Ag, and Au do not depart greatly from those obtained by the thermionic method.

While numerous attempts have been made to measure the threshold frequency for the alkali and alkaline earth metals, none of the results obtained can be considered as very satisfactory. A. F. A. Young's[165] determinations of Φ_0 by photoelectric method, made simultaneously with emission observations, showed that the photoelectric values were consistently higher than those obtained from thermionic data. Undoubtedly even slight traces of residual gases affect the results considerably.

The same remarks apply to a large number of other determinations of threshold frequency for various elements. For this reason discussion of these observations has been omitted. Summarizing the conclusions to be derived from the data in Table XXVII it has been shown by DuBridge that the determination of the work function for emission from the threshold frequency gives reliable results, if care is taken to work with thoroughly degassed surfaces. On the other hand, since it is extremely difficult to maintain a surface in this condition at ordinary temperatures, the photoelectric method is apt to lead to erroneous or at least only approximately correct results.

IX. Contact Difference of Potential

The contact difference of potantial (ΔV) between two surfaces 1 and 2 is connected with the work functions for the two surfaces by the relation,[166]

$$\Delta V = \Phi_1 - \Phi_2 + \frac{T}{11600} ln \frac{A_2}{A_1} \tag{29}$$

where A_1 and A_2 are the values of the constants in the T^2 equations. Usually it has been assumed that $A_1 = A_2$, and that, therefore, the observed contact difference should correspond to the difference in the work functions as measured thermionically. The earlier work has been discussed by Richardson,[166]

[160] R. Suhrmann, Zeits. f. Physik **33**, 63 (1925).
[161] S. C. Roy, Proc. Roy. Soc. **A112**, 599 (1926).
[162] O. W. Richardson, Phil. Mag. **23**, 615 (1912); **24**, 570 (1912).
[163] W. Wilson, Proc. Roy. Soc. **A93**, 359 (1912).
[164] A. Becker, Ann. d. Physik **78**, 83 (1919).
[165] A. F. A. Young, Proc. Roy. Soc. **104**, 611 (1923).
[166] Richardson, Emission of Electricity from Hot Bodies (1921), p. 41.

and further remarks on this topic will also be found in the comprehensive review by K. T. Compton and I. Langmuir,[167] as well as in the treatise by Schottky, Rothe and Simon.

An attempt to check the validity of the above equation by measurements with adsorbed films has been made by I. Langmuir and K. H. Kingdon.[168] The electron current was measured from a standard filament (b), that is, one maintained at a definite temperature and state of activation, to an adjacent filament (a) as a function of the potential difference, V_{ab}. On changing the surface of a, another characteristic was determined and "the voltage displacement of these two characteristics from one another then measured the contact potential difference between the two states of the surface a."

The observed values of ΔV compared with those calculated from values of Φ_0 and A in Table IV are shown in Table XXVIII.

TABLE XXVIII. *Calculated and observed values of the contact potential for various surfaces (Langmuir and Kingdon).*

Surfaces	ΔV (calc.)	ΔV (obs.)
ThW-W	1.81	1.46
CsOW-W*	3.55	3.1
W-OW	4.12	0.8

* The value $A = 0.003$ amp/cm^{2}deg^2 was used for the film.

The comments of the authors on these results are as follows:

"The agreement is not good. In no case was there any indication of a contact potential difference of 4 volts between tungsten and oxidized tungsten. It is believed that the differences between the observed and calculated values are well outside the experimental error in the contact potential measurements for the ThW-W and W-OW combinations. However, it may not be justifiable to use electron emission constants determined at elevated temperatures to calculate contact potential differences between surfaces at room temperature. An attempt was made to measure contact potentials between hot surfaces, but the work was not completed."

It is evident that while the relation between Φ_1, Φ_2 and ΔV has received experimental confirmation in the case of atomically homogeneous surfaces (e.g., pure metals), the same result is not true for composite surfaces. Since, however, the investigations on this problem by J. A. Becker, N. B. Reynolds and W. B. Nottingham, have been fully discussed in the review by Compton and Langmuir, further discussion in the present paper is unnecessary.

X. The Thermionic Constant A

The data on emission constants given in the previous sections show that while the constant A in the T^2 equation should have the theoretical value

$$A_0 \equiv \frac{2\pi mek^2}{h^3} = 60.2 \text{ amps/cm}^2 \text{ deg.}^2 \qquad (30)$$

[167] K. T. Compton and I. Langmuir, Review of Modern Physics, **2**, 124 (1930).
[168] I. Langmuir and K. H. Kingdon, Phys. Rev. **34**, 129 (1929).

actual observations lead to values of A which are both greater and less than A_0. The theoretical value is observed for only a few pure metals; for monatomic films of more electropositive elements, $A < A_0$, while for films of more electronegative elements and for platinum, $A > A_0$.

There have been published a number of papers dealing with the problem of the explanation for this discrepancy, and while it cannot be stated th t a completely satisfactory explanation of all the observations has been obtained nevertheless it is of interest to review these investigations. In the following discussion the subject has been dealt with under the following four headings:

(A) Observations on the relation between A and b_0.

(B) Derivation of A from thermodynamical considerations.

(C) Application of new electron theory to the problem of thermionic emission.

(D) Application of wave mechanics to the same problem.

(A) Observations on the relation between A and b_0.

In 1915, Richardson[169] showed that if the various values of log A for platinum, in vacuum and in presence of gas, are plotted against the corresponding values of b, a *linear relation* is obtained. Similar results were obtained from plots of the emission constants for tungsten as observed by Langmuir in vacuum and in presence of various gases,[170] and also by plotting the emission constants for potassium.[171]

L. A. DuBridge[172] has pointed out more recently that a relation of the form

$$\log_{10} A = c + s \cdot b_0 \tag{31a}$$

apparently holds valid for the emission data from a number of surfaces. In this relation c and s are empirically determined constants.

The same type of relation has also been found to hold valid for emission constants for zirconium and hafnium during ageing, and for oxide coated cathodes during activation.

From the linear plots given by DuBridge, and from data published by other investigators, the values of the constants in equation (31a) may be calculated. Adopting the method used by DuBridge, this relation may be expressed in the form

$$\log A = \log A_0 + \beta(b - b_0),^* \tag{31b}$$

where $\beta = 2.303\ s$ in equation (31a). This may again be expressed, as suggested by Schottky in the form,

$$A = A_0 \epsilon^{\alpha} \tag{32}$$

[169] O. W. Richardson, Proc. Roy. Soc. **A91,** 524 (1915).

[170] Richardson, Emission of Electricity, p. 135 (1921).

[171] Richardson and Young, Proc. Roy. Soc. **A107,** 393 (1925).

[172] L. A. DuBridge, Proc. Nat. Acad. **14,** 788 (1928).

* Unless otherwise specified, natural logs are to be understood in this and subsequent equations.

where $\alpha=\beta(b-b_0)$, and substituting this in the T^2 equation for emission, we have.

$$I = A_0T^2\epsilon^{-(b-\alpha T)/T} = A_0T^2\epsilon^{-b'/T} \quad (33)$$

where $$b'=b-\alpha T \quad (34)$$

In Table XXIX the values of β and b_0 have been calculated by means of equation (31b). The last column gives the value of α corresponding to the values b and Φ in the preceding columns. It will be observed, that in the case Th-W, Cu (liquid), and oxide coated cathodes, where $b<b_0$, α *is negative.*

TABLE XXIX. *Constants in the relation between A and b_0 for various emitters.*

Emitter	Authority	$\beta\times10^4$	$b_0\times10^{-4}$	$b\times10^{-4}$	Φ	ϵ^α
W	Du Bridge	4.56	5.24	6.40	5.53	1.98×10
Th-W	Du Bridge	1.76	5.24	3.05	2.63	2.11×10^{-2}
Pt	Du Bridge	3.41	5.88	7.27	6.27	1.14×10^{2}
K	Du Bridge	24.70	0.748	1.908	2.21	2.75×10^{12}
Cu (liquid)	Du Bridge	6.91	5.27	4.11	3.54	3.31×10^{-4}
Zr, Hf	Zwikker	4.08	4.46	5.62	4.84	1.13×10^{2}
BaO	Detels	6.79	2.87	1.71	1.47	3.78×10^{-4}

Further reference to this table will be made in a subsequent section, where the significance of both equation (34) and the coefficient ϵ^α will be discussed more fully.

(B) Thermodynamical derivation of A.

Richardson's derivation of an equation for emission was based on the classical electron theory of metallic conduction. Assuming that the free electrons within the metal are held within the metal by an electric force at the surface and that only those electrons can escape which possess energy in excess of a certain critical value, he showed that the electron emission as a function of the temperature must be expressed by an equation of the form[173]

$$I = aT^{1/2}\epsilon^{-b/T} \quad (35)$$

where $a=N(kT/2\pi m)^{1/2}$
and $N=$number of electrons per cm^3 of the metal
$k=$Boltzmann constant,
$m=$mass of electron.
Subsequently Richardson pointed out[174] that an equation of the form

$$I = AT^2\epsilon^{-b_0/T} \quad (2)$$

where A is a universal constant, could be derived on the basis of the quantum theory. In the derivation of this constant, he applied certain hypotheses

[173] O. W. Richardson's Treatise, also given by I. Langmuir, reference (49).
[174] O. W. Richardson, Phil. Mag. **28**, 633 (1914).

which had been utilized by Debye and Keesom in calculating specific heats at low temperatures, and obtained a relation of the form

$$A = (k^2me/h^3)f,$$

where e denotes the electronic charge and f corresponds to a numerical coefficient different from 2π, so that the actual value calculated from his relation corresponds to about 5 amp. cm^{-2} $deg.^{-2}$ instead of 60.2 as calculated subsequently by Dushman.

M. v. Laue[175] and W. Schottky[176] showed that the electrons emitted from an incandescent solid could be regarded as an electron gas having the properties of an ideal monatomic gas, and R. C. Tolman[177] applied the third law of thermodynamics to calculate the entropy of this electron gas.

It was these considerations that led Dushman[178] to the deduction that the constant A in equation (2) should be a universal constant of which the value is determined, in accordance with the theory derived by Sackur and Tetrode, by the third law of thermodynamics, with the value A_0 given by (30).

The considerations on which this result was based may be presented briefly as follows:

The emission of electrons from a metal may be regarded as thermodynamically equivalent to the evaporation of a monatomic gas of molecular weight Nm, where N is Avogadro's number ($=6.06\times10^{23}$) and m is the mass of the electron. That is, at any temperature, T, there exists an equilibrium pressure, P, of the electron gas, and in accordance with the second law of thermodynamics the latent heat of evaporation at this temperature is given by Clapeyron's equation

$$L = RT^2\frac{d\log P}{dT} \tag{36}$$

where L = heat of vaporaization per gram molecular weight
and R = gas constant per gram molecular weight
$= Nk = 1.987$ cal./deg.

Hence

$$\log P = \frac{1}{R}\int^T \frac{LdT}{T^2} + i \tag{37}$$

where i is an integration constant.

If C_p denotes the specific heat of the electron gas at constant pressure, and c_p that of the electrons in the metal, L can be expressed as a function of T by the relation

$$L = L_0 + \int_0^T C_p dT - \int_0^T c_p dT \tag{38a}$$

[175] M. v. Laue, Jahrb. d. Rad. **15**, 205, 257 (1918); Ann. d. Physik **58**, 695 (1919) Phys. Zeits. **20**, 202 (1919).

[176] W. Schottky, Phys. Zeits. **15**, 872 (1914); **20**, 49, 220 (1919).

[177] R. C. Tolman, J. Am. Chem. Soc. **42**, 1185 (1920); **43**, 866 (1921).

[178] S. Dushman, Phys. Rev. **21**, 623 (1923).

where L_0 = heat of vaporization at $T = 0$.

For a monatomic gas, $C_p = (5/2)R$; while for electrons in a solid the experimental evidence shows that as a first approximation we may assume $c_p = 0$. Hence (38a) becomes

$$L = L_0 + 5RT/2 \tag{38b}$$

O. Sackur[179] and H. Tetrode[180] showed that in the case of a monatomic gas

$$i = \log\left[\frac{(2\pi)^{3/2}k^{5/2}\cdot m^{3/2}}{h^3}\right]. \tag{39}$$

Introducing these relations into (37) it follows that

$$P = \frac{(2\pi)^{3/2}k^{5/2}m^{3/2}}{h^3}T^{5/2}\epsilon^{-L_0/RT}. \tag{40}$$

But according to the kinetic theory of gases, the number of electrons, n, striking the cathode per unit area in unit time, is given by the relation

$$n = P/(2\pi mkT)^{1/2} \tag{41}$$

Therefore, if we assume no reflection of electrons, the current per unit area is given by

$$\begin{aligned} I &= ne = Pe/(2\pi mkT)^{1/2} \\ &= A_0T^2\epsilon^{-b_0/T} \end{aligned} \tag{2}$$

where $b_0 = L_0/R$.

A more rigorous investigation has been made by P. W. Bridgman[181] of the conditions under which the coefficient in equation (2), which we shall denote by A, is equal to A_0. The conclusion is reached that "A necessary and sufficient condition is that the difference between the specific heat of neutral metal and of a charge on its surface vanish," that is,

$$C_{pm} - C_{p\rho} = 0 \tag{42}$$

where the subscripts m and ρ refer to metal and surface charge respectively.

Theoretically it is necessary, of course, to take into account the fact that the electrons in evaporating leave a positive charge on the surface. This was first pointed out by Bridgman[182] and in the second paper, referred to above, he shows that while condition (42) is probably valid for uncoated metals, it may not hold for coated surfaces, or the entropy of the surface charge may not vanish at $T = 0$.

The problem has also been investigated by L. Tonks and I. Langmuir,[183] and the following discussion summarizes their arguments.

[179] O. Sackur, Ann. d. Physik **36**, 598 (1911); **40**, 67 (1913).
[180] H. Tetrode, Ann. d. Physik **38**, 434 (1912).
[181] P. Bridgman, Phys. Rev. **27**, 173 (1926).
[182] P. Bridgman, Phys. Rev. **14**, 306 (1919).
[183] L. Tonks and I. Langmuir, Phys. Rev. **29**, 524 (1927); L. Tonks, **32**, 284 (1928).

The term, L, in equation (36) is actually made up of two terms, that is,

$$L = L_\rho + L_s \tag{43}$$

where L_ρ = cooling effects due to evaporation of electrons under reversible conditions,
and L_s = latent heat of surface charge.

If L_c denote the cooling effect measured calorimetrically, then

$$L_\rho = L_c + kT/2. \tag{44}$$

Furthermore, from emission data there is obtained the derivative

$$-k\frac{d\log(I/T^2)}{d(1/T)} = -kb' \tag{45}$$

where b' corresponds to the quantity previously designated by b_0.

From these relations it follows, as shown by Tonks and Langmuir, that

$$L_s = kb' - L_c + 2kT. \tag{46}$$

Now according to the measurements of Davisson and Germer on tungsten,[184] $kb'/e = -4.48$ volts, $L_c/e = 4.91$ volts, and for $T = 2270$, $2kT/e = -0.391$ volts. Hence $L_s/e = 0.04$ volt, which is less than the probable error of the measurements. It therefore appears "that the surface heat of charging for all pure metals is probably but a small fraction of the work function."

This conclusion is also in agreement with Millikan's observations on stopping potentials for clean surfaces of the alkali metals in vacuum. From equations (27) and (29) it follows that the contact potential between surfaces 1 and 2 is given by

$$\begin{aligned} eV_{12} &= h(\nu_{01} - \nu_{02}) \\ &= kT\log(A_2/A_1) + k(b_1 - b_2) \end{aligned} \tag{47}$$

Since $h\nu_0 = kb$, it follows that A must be the same for all metals.

For the difference in L_s for two surfaces, the relation derived is of the form

$$L_{s1} - L_{s2} = kT\log(A_1/A_2) + eP_{12} \tag{48}$$

where P_{12} is the Peltier heat. As this is negligible, it is possible to calculate the following values of L_s/e from the values of A given previously, assuming $L_s = 0$ for tungsten:

Th-W	-0.39 V at $T = 1500°$K
Cs-O-W	-0.62 V at $T = 650°$K
O-W	$+2.92$ V at $T = 1600°$K

The signs show that charging the surface positive is attended by the evolution of heat in the first two cases, and by absorption in the last case. The values of L_s calculated by means of equation (48) are evidently considerable, but since no direct measurements of L_c are available, it is not possible to check these conclusions.

[184] Davisson and Germer, Phys. Rev. **20**, 300 (1922), see Table 12.

It should be noted that according to Bridgman's critical examination of the problem, "It is not necessary that the surface heat vanish in order that the emission formula (2) shall be valid; provided only that it be proportional to the temperature."

In a still more recent paper on the photoelectric effect,[185] Bridgman has deduced the conclusion that the value of A must depend not only on that of L_s but also on the variation in ν_0 with temperature.

The difference between observed values of A and the theoretical value A_0 has been ascribed by K. F. Herzfeld[186] as due, at least in part, to the energy of expansion of the emitting substance. As a consequence Φ_0 must depend on the specific volume, and the observed value of A must be greater than A_0, whenever the temperature coefficient of expansion is positive. Furthermore, this increase in A must be accompanied by a shift to the red in the threshold frequency, ν_0, with increase in temperature. DuBridge actually observed a shift of about 50A between 800° and 1200°C. This corresponds to a value for α in equations (32) and (34) of 4.5, and would, therefore, give a value of $\epsilon^\alpha = A/A_0 = 90$, whereas the observed ratio is approximately 250. On the other hand, in case of tungsten, the value of ν_0 shows no change with temperature, and therefore we should have $\alpha = 0$ and $A = A_0$. Herzfeld's theory thus gives the right order of magnitude for the ratio A/A_0 in at least two cases. Whether it is capable of accounting for negative values of α is, however, questionable. It would be of interest in this connection to investigate more thoroughly the thermionic emission from the alkali metals and also the effect of temperature on the values of the photoelectric threshold frequencies.

(C) Thermionic emission on basis of new electron theory.

So far the derivation of the constant A has been considered from the point of view of classical thermodynamics. In the present section we shall discuss a treatment of the problem of thermionic emission, on the basis of the statistical mechanics of Fermi and Dirac, which we owe to A. Sommerfeld[187] and other investigators associated with him.

The starting point for Sommerfeld's theory is the formula derived by Fermi[188] for the distribution-in-energy of an assemblage of N particles in volume V at temperature T. This may be written in the form

$$F(E) = \frac{2\pi G V (2m)^{3/2}}{h^3} \cdot \frac{E^{1/2}}{B^{-1}\epsilon^{E/kT} + 1} \tag{49}$$

where G = the statistical weight
m = mass of particle.

[185] P. Bridgman, Phys. Rev. **31**, 90 (1928).

[186] K. F. Herzfeld, Phys. Rev. **35**, 248 (1930).

[187] A. Sommerfeld, Zeits. f. Physik **47**, 1 (1928).

[188] The remarks on this theory in the present section are based largely on the comprehensive discussion of the subject by K. K. Darrow, Phys. Rev. Supplement **1**, 90 (1929). See also the recent treatise by L. Brillouin, "Les Statistiques Quantiques," Paris (1930).

$$N = \int_0^\infty F(E)dE \tag{50}$$

while B is a function defined as follows:

(1) For a gas at high temperatures or low pressures, the Maxwell-Boltzmann distribution law is known to apply very satisfactorily. Hence B must be so small that the second term in the denominator of (49) is negligible by comparison with the first. Since in the classical statistical mechanics the formula derived is of the form

$$F(E) = \frac{2N\pi}{(\pi kT)^{3/2}} E^{1/2}\epsilon^{-E/kT} \tag{51}$$

it follows that for this case

$$B = \frac{Nh^3}{GV(2\pi mkT)^{3/2}}. \tag{52}$$

Let us consider the case of 1 gm molecular weight of hydrogen ($m = 3.33 \times 10^{-24}$) at $T = 10$ and V(molecular volume) = 2,240 cm.^{3} Assuming $G = 1$, $B = 1.56 \times 10^{-3}$. Since B decreases with increase in m, V and T, it is evident that in case of gases, even at high pressures, B is extremely small compared to 1, and that consequently $B^{-1}\epsilon^{E/kT}$ will be very large compared to unity.

(2) The conditions are, however, quite different for the electrons in a metal. Assuming that the concentration of electrons is the same as that of the atoms it is found that B is very large, even at 1000–2000°K. Thus for tungsten at ordinary temperatures, $N/V = 6.36 \times 10^{22}$. Substituting $m = 9 \times 10^{-28}$ and $T = 2000$ in equation (51), and assuming $G = 2$, we obtain the result $B = 146$ approximately. This indicates that equation (52) cannot be used to substitute for B in equation (49). The gas is said to be in a "degenerate" state.

Carrying out the integration indicated by (50), Fermi has obtained, as a first approximation, the result

$$2mkT \log B = h^2 \left(\frac{3n}{4\pi G}\right)^{2/3} \tag{53}$$

where $n = N/V$ = number of corpuscles per cm.^{3}

Equation (53) can be written in the form.

$$B^{-1} = \epsilon^{-W_i/(kT)} \tag{54}$$

where

$$W_i = \frac{h^2}{2m}\left(\frac{3n\nu}{4\pi G}\right)^{2/3}$$

n = number of atoms per cm.^{3} = N/V
V = atomic volume
$N = 6.062 \times 10^{23}$
ν = number of free electrons per atom (assumed to be the same as the number of valence electrons)
$G = 2$ for electron.

The magnitude W_i which has been introduced here is evidently a quantity of energy and may be expressed in terms of atomic constants in the following forms:

$$\left.\begin{aligned} W_i &= 5.78 \times 10^{-27}(n\nu)^{2/3} \text{ ergs} \\ &= 3.634 \times 10^{-15}(n\nu)^{2/3} \text{ electron-volts} \\ &= 26.07\nu^{2/3}V^{-2/3} \text{ electron-volts} \end{aligned}\right\} \quad (55)$$

For tungsten and potassium, *assuming* $\nu = 1$, the values thus calculated for W_i are 5.77 and 2.1 volts respectively. Sommerfeld designates W_i as the "inner work function." It represents the zero point energy due to the kinetic energy of agitation of the electrons in the metal. It thus corresponds to the thermodynamical function, ψ, defined by Gibbs as the *free energy at constant pressure* (per electron). The electron pressure may be derived from this quantity in accordance with the relation

$$p_0 = 2W_i n\nu/5 \quad (56)$$

which gives values for p_0 of the order of 2×10^5 atmospheres approximately.

Substituting in equation (49) from (54) it is seen that the exponential term in the denominator becomes

$$B^{-1}\epsilon^{E/(kT)} = \epsilon^{(E-W_i)/(kT)} \quad (57)$$

which leads to the following conclusions regarding the form of $F(E)$ for $T = 0$.

For $E < W_i$ the exponential in (57) becomes equal to zero. Hence $F(E) = CE^{1/2}$ where C denotes the coefficient in equation (47).

For $E > W_i$, the exponential becomes infinitely great. Hence $F(E) = 0$.

Thus the form of $F(E)$ at $T = 0$ is that of the full curve shown in Fig. 20. As the temperature is increased from $T = 0$ the plot for $F(E)$ assumes the form shown in the dotted curve.

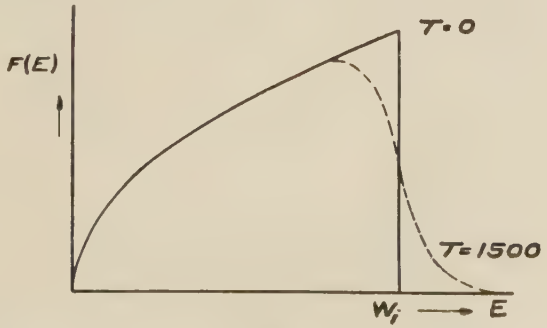


Fig. 20. Distribution of energy among electrons in a metal by the Fermi statistics.

The corresponding plots for the distribution in velocities $u = (\dot{x}^2 + \dot{y}^2 + \dot{z}^2)^{1/2}$ (the velocities in random directions) are shown in Fig. 21. The velocity u_0 is given by the relation

$$\tfrac{1}{2}mu_0^2 = W_i. \quad (58)$$

The dotted curve gives the plot for $T = 1500°K$. It shows that for velocities in excess of u_0 a practically Maxwellian distribution is obtained.

From these distribution functions it is possible, as shown by Sommerfeld, to deduce the equation for emission by the following procedure. It is assumed

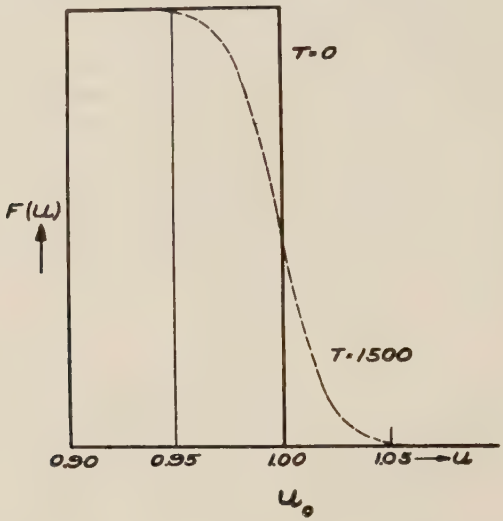


Fig. 21. Distribution of velocities among the electrons in a metal by the Fermi statistics.

that the thermionic current is produced by all those electrons which strike the boundary surface of the metal with velocities such that the *components normal to the surface*, $\dot{x}$, are greater than those corresponding to the relations

$$\tfrac{1}{2}mx_0^2 = W_a \tag{59}$$

where W_a is the "external work function."

Denoting by $N(W)\,dW$ the number of electrons having *normal energies* between W and $W+dW$ which strike unit area per unit time, the relation deduced is of the form[189]

$$N(W) = \frac{4\pi m}{h^3} kT \log\left(1 + \epsilon^{-(W-W_i)/(kT)}\right) \tag{60}$$

Hence

$$\left.\begin{aligned} N(W) &= \frac{4\pi m}{h^3}(W_i - W) \text{ for } \frac{W - W_i}{kT} \ll 0 \\ &= \frac{4\pi m}{h^3} kT \text{ (approx.) for } W \leqq W_i \\ &= \frac{4\pi m}{h^3} \cdot kT\epsilon^{-(W-W_i)/kT} \text{ for } \frac{W - W_i}{kT} \gg 0 \end{aligned}\right\} \tag{61}$$

Consequently the distribution curve is linear for $T = 0$, as shown in Fig. 22, and takes the form of the dotted curve for $T \gg 0$. For values of $W > W_i$,

[189] The derivation here given is taken from the paper by L. Nordheim, Phys. Zeits. **30**, 177 (1929).

this part of the curve follows a Maxwellian distribution law, which is in agreement with observations on the velocity distribution for the electrons emitted from a hot cathode.

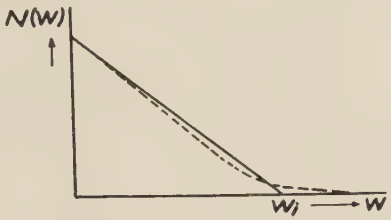


Fig. 22. Distribution function for electrons for energy normal to surface.

By forming the integral of $N(W)dW$ for the limits $\dot{x}=\dot{x}_0$ to $\dot{x}=\infty$, and for the limits of $\dot{y}$ and $\dot{z}$ the values 0 and ∞, the electron current per unit area is obtained in the form

$$I = A_0 G T^2 \epsilon^{-(W_a - W_i)/kT} \tag{68}$$

It should be observed that if the "classical" value of B, given by equation (52), is used, the resulting equation for emission is that derived originally by Richardson (equation (35)) in which W_a/k corresponds to his value b.

Equation (62) differs from that derived on the basis of classical thermodynamics in two respects:

(1) The presence of the factor G, to which has been assigned the value 2 by R. H. Fowler, on the basis of the Goudsmit-Uhlenbeck hypothesis of electron spin,

(2) The substitution in the exponential factor of the classical b_0 by $(W_a - W_i)/k$.

However, the fact that both classical thermodynamics and the Fermi statistics lead to the same value A_0 for the coefficient of T^2, is extremely significant. Apparently the statistical method does not indicate any reason for deviations in the observed values of A from a theoretical value $2A_0$. Before proceeding to discuss the views which Nordheim and Fowler have advanced to account for such deviations, it is, however, necessary to make a few remarks about the energy quantity, W_a, introduced in the above equations.

This energy is the work required to overcome the attraction of the positive charges in the metal for the electrons. It corresponds, therefore, to the energy required to remove an electron (initially at rest) from the crystal lattice, and may be identified, according to H. Bethe[190] and L. Rosenfeld and E. E. Witmer,[191] with the potential energy of the electrons in the lattice as calculated from the refractive index, μ, of the crystal for electron beams. If V denote the volt-velocity of the electrons in the incident beam, the "wave-length" of the electrons in vacuum is given by the de Broglie relation,

[190] H. Bethe, Naturwiss. **16**, 333 (1928).

[191] L. Rosenfeld and E. E. Witmer, Zeits. f. Physik **49**, 534 (1928).

$$\lambda = h/(2meV)^{1/2} \tag{62}$$

and the refractive index, by

$$\mu = (1 + W_a/V)^{1/2} \tag{63}$$

where W_a is naturally expressed in electron-volts.

Values of W_a calculated by the investigators mentioned above range between 10 and 20 volts, while W_i, as calculated by equation (54), on the assumption that $\nu = 2$ or 3, is about 2 to 6 volts less than W_a. This is in agreement with the values of

$$b_0 = \frac{W_a - W_i}{k} \tag{64}$$

determined from emission data.

(D) Theory of penetration of electrons through potential barrier.

According to Sommerfeld's derivation of the equation for emission, it is assumed that *all* the electrons in the metal reaching the inner wall with normal energy greater than W_a are transmitted. However, L. Nordheim[192] has shown, from considerations based on the application of wave mechanics, that in general some of the electrons will be reflected at the boundary. The transmission coefficient, D, is a function of the normal kinetic energy W, denoted by $D(W)$, and hence the actual emission is given in any particular case by the relation

$$I = A_0 G \bar{D} T^2 \epsilon^{-(W_a - W_i)/(kT)} \tag{65}$$

where $\bar{D}$ is a mean transmission coefficient.

The value of $\bar{D}$ depends upon the nature of the potential distribution on both sides of the boundary. It may be calculated for any given distri-

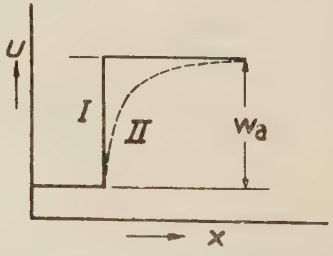


Fig. 23. Potential energy as function of distance from homogeneous cathode.

bution by considering the transmission of electrons across the boundary as "the passage of plane wave-trains of electrons over (or through) a hill of potential energy."[193]

Let us consider, for instance, a potential distribution such as that shown in Fig. 23 by the straight line plot (I). For $0 < x_0$, the potential energy $U = 0$,

[192] L. Nordheim, Zeits. f. Physk. **46**, 833 (1927); Physikal. Zs. **30**, 177 (1929); Proc. Roy. Soc. **A121**, 626 (1928).

while for $x>0$, $U=W_a$. Owing to the effect of the image force, the potential energy for $x>0$ will have the value $W_a-e^2/4x$ and will therefore be represented by the dotted curve II. Solving the Schroedinger differential equation in ψ for these conditions, Nordheim finds that for case I, Fig. 23,

$$\left.\begin{aligned} D(W) &= 0 \text{ for } W < W_a. \\ D(W) &= \frac{4[(W-W_a)W]^{1/2}}{[W^{1/2}+(W-W_a)^{1/2}]^{1/2}} \text{ for } W > W_a \end{aligned}\right\} \tag{66}$$

The *mean* value in the latter case is given by

$$\bar{D} = 2(\pi kT/W_a)^{1/2}. \tag{67}$$

This means, according to Fowler,[193] that, "The resulting current at the ordinary temperatures of thermionic emission is very much the same as if 20 percent of all those electrons are transmitted which are incident on the boundary with energy greater than W_a for more gradual slopes the fraction is higher and rapidly approaches unity as the slope is eased off. For the image force hill of case II, Fig. 23, the fraction is about 95 percent." Hence, wave mechanics makes no serious modifications to classical calculations, and D is approximately unity.

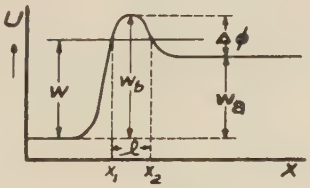


Fig. 34. Potential energy as function of distance for electropositive surface layer on cathode.

It may, however, happen that the potential energy as a function of x is represented by a "hill" such as shown in Fig. 24, in which the potential energy which the electrons have to exceed in order to cross the boundary is *diminished* from its normal value, W_b, to a *lower value*, W_a, by a surface layer (as, for instance, in the case of films of thorium on tungsten).

According to classical mechanics, an electron having the kinetic energy W, where, $W_b>W>W_a$, could not penetrate the barrier; that is, $D(W)=0$ for $W<W_b$. But *in the new mechanics there exists a certain definite probability that the electron wave-train can pass through the hill.* The probability increases with decrease in the width, l, of the hill, and is significant for values of l which correspond to only a few de Broglie wave-lengths. The analogy with a similar phenomenon in optics has been pointed out by various writers, while the principle here stated has also been used by R. W. Gurney and E. U. Condon[194] and by G. Gamow[195] to develop a theory of radioactive disintegration.

[193] R. H. Fowler, Robert Boyle lecture, Oxford, May 1929. This will be referred to as "R. B. Lect."

[194] R. W. Gurney and E. U. Condon, Phys. Rev. **33**, 127 (1929).

The probability of transmission of the electrons varies exponentially with the values of l and of $\Delta\Phi = W_b - W_a$. Nordheim[196] has calculated $D(W)$ for the idealized case shown in Fig. 25, with the results shown in the plot in Fig. 26.

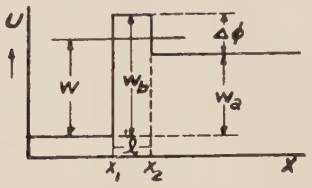


Fig. 25. Simplified potential energy curve for same case as in Fig. 24.

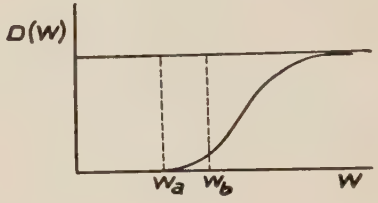


Fig. 26. Transmission coefficient for electrons for case shown in Fig. 25.

For $W > W_b$, $D(W)$ is approximately unity. For $W_b > W > W_a$ and the condition

$$\epsilon^{-(W_b - W_a)/(kT)} \ll \epsilon^{-2\sigma l (W_b - W_a)^{1/2}} \tag{68}$$

$$\bar{D} \sim \frac{8(\pi k T W_a)^{1/2}}{W_b} \epsilon^{-2\sigma l (W_b - W_a)^{1/2}} \tag{69}$$

where $\sigma = 8\pi^2 m/h^2 = 4.069 \times 10^{13}$ cgs units
and l = width of potential barrier.

Substituting for $\bar{D}$ in equation (65) it is evident that

$$\begin{aligned} A &= A_0 G \bar{D} \\ &= A_0 G \frac{8(\pi k T W_a)^{1/2}}{W_b} \epsilon^{-2\sigma l (W_b - W_a)^{1/2}} \end{aligned} \tag{70}$$

Hence

$$\log A = \log C T^{1/2} - 2\sigma l (k \cdot \Delta b_0)^{1/2} \tag{71}$$

where C is a constant, and $k \cdot \Delta b_0$ *is the decrease in work function due to the surface layer of thickness l.*

It will be recognized that this relation is similar in form to the empirical relation, equation (31a) or (31b), deduced by DuBridge and thus *accounts for the decrease in both A and b_0 observed for electropositive films.*

For instance, let us consider the case of Th-W for which $A = 3.0$, $b_0 = 30$, $500 = 2.63$ volt/k, while for tungsten $A = 60.2$, $b_0 = 52{,}400 = 4.53$ volt/k. Assuming the theoretical value $A_0 G = 120$, it follows that for Th-W, $\bar{D} = 0.025$.

$$W_a = 5.77 + 2.63 \text{ volts} = 8.40 \times 1.59 \times 10^{-12} \text{ erg.}$$

$$W_b = 5.77 + 4.53 \text{ volts} = 10.30 \times 1.59 \times 10^{-12} \text{ ergs.}$$

---

[195] G. Gamow, Zeits. f. Phys. **51,** 204 (1928).

[196] See also R. H. Fowler, Proc. Roy. Soc. **A122,** 36 (1929) for a correction of certain formulae derived by Nordheim, reference 192.

For $T = 1500°K$

$$\bar{D} = 0.025 = 1.4\epsilon^{-2\sigma l .106(1.9 \times 1.59)^{1/2}}$$

Therefore $l = 2.85 \times 10^{-8}$ cm. This appears to be a perfectly reasonable value for the thickness of a monatomic film. It can also be shown that this satisfies the condition expressed by relation (68).

From equations (71) and (31b) it follows that we should find

$$2\sigma l(k \cdot \Delta b_0)^{1/2} = \beta \Delta b_0$$

or

$$l = \frac{\beta \times 10^{-5}}{9.53}(\Delta b_0)^{1/2}.$$

For $\Delta b_0 = 10{,}000$, and the values of β given in Table XXIX, this relation leads to values of l varying from 10^{-7} to 10^{-8} cm, which again seems a reasonable, result.

While this theory apparently accounts for values of A less than $2A_0$, it is inadequate to explain the results obtained by DuBridge for clean platinum and by Kingdon for oxygen on tungsten. With regard to the latter, Fowler remarks as follows:[197] "It is impossible to believe here that $\Phi_0 (= 9.2$ volts) is a true work function and $A (= 5 \times 10^{11})$ an emission coefficient. It is much more likely that a great part of the whole new factor $10^{10}\ \epsilon^{-4.7/kT}$ introduced by the oxygen is really a temperature dependent emission coefficient, representable in this form over a short temperature range."

Schottky has suggested, as mentioned previously, that

$$A = A_0\epsilon^{\alpha} \tag{32}$$

where

$$b = b_0 - \alpha T. \tag{34}$$

Hence, if these relations are to account for DuBridge's relation (31a) it is necessary, as in (31b) to assume that α varies linearly with changes in b_0. Fowler has discussed this assumption, which amounts essentially to the hypothesis of a *negative* specific heat of electricity for platinum and oxygen, and concludes that this theory is not admissible.

Perhaps it is best, under the circumstances, to await further experimental data before attempting to develop any theoretical explanations. Furthermore, if Nottingham's conclusion is valid that by correcting for the effect of field strength, the value of A for Th-W is increased to practically the theoretical value A_0, then it may prove that the theory discussed in the previous paragraphs is not required. It may be that further investigations on other emitters for which apparently $A < A_0$ will show that this is due to incorrect extrapolation to zero field strength, while the explanation offered by Herzfeld for platinum and by Fowler for oxygen may be found adequate for these cases.

[197] Fowler, Proc. Roy. Soc. **A122,** 36 (1929).

XI. Electron Emission in Intense Electric Fields

The investigations on the emission of electrons in presence of very intense fields have been discussed in the review by Compton and Langmuir. It has been shown by the investigations of Millikan and Eyring;[198] Millikan and Lauritsen;[199] de Bruyne;[200] and the research staff of the General Electric Company, London,[201] that with extremely high anode voltages, emission occurs even at room temperature, and that the observations on this so-called "cold cathode effect" or "auto-electronic" emission are in agreement with the empirical relation,

$$I = A(T + cF)^2 \epsilon^{-b/(T+cF)} \tag{72}$$

where F is the field strength and A and c are constants which vary with the condition of the surface.

An explanation of this phenomenon has been deduced on the basis of Schroedinger's theory by J. R. Oppenheimer[202] and by R. H. Fowler and L. Nordheim.[203] In the following discussion, we shall summarize the derivation given by the last named investigators, which follows the same line of reasoning as was involved in the consideration of the value of the emission constant A in the previous section.[204]

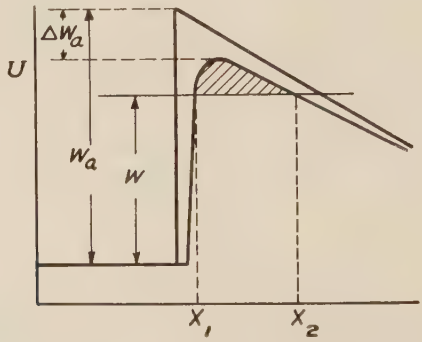


Fig. 27. Variation in potential energy with distance for case of intense electric fields.

For relatively weak fields, the effect of field strength, as mentioned previously, is given by Schottky's theory. The decrease in work-function varies with field strength in accordance with equation (8) which may be stated in the form

$$\Delta W_a = eF^{1/2}. \tag{73}$$

[198] Millikan and Eyring, Phys. Rev. **27,** 51 (1926); **31,** 900 (1928).

[199] Millikan and Lauritsen, Proc. Nat. Acad. Sci., **14,** 15 (1928); Phys. Rev. **33,** 598 (1929).

[200] de Bruyne, Proc. Camb. Phil. Soc. **120,** 423 (1928).

[201] Phil. Mag. **1,** 609 (1926).

[202] J. R. Oppenheimer, Phys. Rev. **31,** 66 (1928); Proc. Nat. Acad. Sci., **14,** 789 (1928).

[203] R. H. Fowler and L. Nordheim, Proc. Roy. Soc. **A119,** 173 (1928); L. Nordheim, Physikal Zs. **30,** 177 (1929); Fowler, R. B. Lecture.

[204] A solution of the problem has also been given by C. Eckart, Phys. Rev. **35,** 1303 (1930).

The value of F is determined by the *geometrical* arrangement and dimensions of the cathode. In presence, however, of very strong fields the emission occurs from points or extremely small areas and is due to the fact that according to the new mechanics there exists a definite probability that electrons can traverse a potential barrier.

Under these conditions the potential distribution at an emitting spot is of the form shown in Fig. 27. Owing to the Schottky effect, the sharp corner indicated in the upper plot becomes rounded off as indicated in the lower curve, the decrease ΔW_a being given by equations (73). Because the temperature is low, there are no incident electrons with sufficient energy to go over the top of the hill, and therefore *all must go through the barrier.*

The potential energy, U, is of the form

$$\left.\begin{aligned} U &= W_a - \frac{e^2}{4x^2} - xF \quad \text{for } x > x_0 \\ &= 0 \qquad\qquad\qquad\quad \text{for } x < x_0 \end{aligned}\right\} \tag{74}$$

with

$$U_{\max} = W_a - \Delta W_a. \tag{75}$$

The solution of Schroedinger's equation for these conditions yields for the electron emission at $T = 0$ the relation

$$I = \frac{e}{2\pi h W_a}\left(\frac{W_i}{b_0 k}\right)^{1/2} F^2 \epsilon^{-\alpha/F} \tag{76}$$

where

$$\alpha = \frac{4}{3}(b_0 k)^{3/2}\left(\frac{8\pi^2 m}{h^2}\right)^{1/2}$$

and W_i, W_a and $b_0 k$ express the inner work function, external work function and ordinary thermionic work function respectively in the same cgs units.

For these energies in *volts*, I in *amps/cm*^{2} and F in volts/cm,[205]

$$I = \frac{6.2 \times 10^{-6}}{W_a}\left(\frac{W_i}{\Phi_0}\right)^{1/2} \epsilon^{-6.8 \times 10^7 \Phi_0^{3/2}/F} \tag{77}$$

Hence at low temperatures the electron current should be given by a relation of the form

$$I = BF^2 \epsilon^{-\alpha/F} \tag{78}$$

where α and B are constants, or $\log(I/F^2)$ plotted against $1/F$ should yield a straight line. As in the case of temperature emission, the effect of the term F^2 is so small compared to the exponential term that experimentally this also amounts to a linear relation between $\log I$ and $1/F$, a result which is in agreement with the empirical relation expressed in equation (72).

Equation (77) has been subjected to a more rigorous experimental test in an extensive investigation by Stern, Gossling and Fowler.[205] As they

[205] The numerical factor in the exponent is the corrected value given in a subsequent paper by T. E. Stern, B. S. Gossling and R. H. Fowler, Proc. Roy. Soc. **A124**, 699 (1929).

point out, both the emitting area, S, and the effective field-strength, $F=\beta F_m$ (where β is a constant and F_m the measured field strength) may be derived by means of relations deduced from equation (76) or (77) From data obtained by previous investigators, especially by Gossling,[2] [6] and Millikan and Eyring,[198] the value of β is found to vary from 1.5 to 115, with values of S ranging from 1.2×10^{-10} cm^2 to 8×10^{-8} cm^2 in different runs. The test of the validity of equation (77) is the fact that the calculated value of S comes out a constant for a series of values of F in any one set of observations under the same conditions.

Of special interest is the discussion by the above mentioned investigators of the effect of a surface-layer in reducing simultaneously both B and α in equation (78), the reduction in α corresponding to a decrease in Φ_0 for the emitting surface because of the presence of a surface film.

A solution of the Schroedinger equation for this case leads to the deduction that the film-thickness (denoted by a) increases as Φ_0 decreases. If the emitting area is calculated on the assumption of a film-free surface (as in equation (76)) a value of S is derived which is considerably less than the true emitting area S_0 derived for the case of a film of definite thickness. Table XXX shows values of these different quantities calculated from three different experimental sets of observations. The first horizontal line for each set of observations refers to the film-free surface. These observations were

TABLE XXX. *Film thickness and emitting area of cathodes in intense electric fields.*

Number of experiment	β	Φ_0	S(cm^2)	S_0(cm^2)	a(cm)
	19.4	4.5	5.1×10^{-9}	5.1×10^{-9}	
1	19.4	2.81	4.6×10^{-10}	5.1×10^{-9}	1.5×10^{-8}
	19.4	1.58	1.8×10^{-10}	5.1×10^{-9}	2.5×10^{-8}
	6.72	4.5	1.2×10^{-10}	1.2×10^{-10}	
2	6.72	1.41	4.8×10^{-13}	1.2×10^{-10}	3.9×10^{-8}
	6.72	1.50	1.5×10^{-10}	1.2×10^{-10}	0.5×10^{-8}
	6.72	1.52	8.7×10^{-12}	1.2×10^{-10}	2.2×10^{-8}
	2.9	4.5	8×10^{-8}	8×10^{-8}	
3	2.9	3.95	3.9×10^{-8}	8×10^{-8}	0.4×10^{-8}
	2.9	2.85	4.1×10^{-9}	8×10^{-8}	1.8×10^{-8}
	2.9	2.33	8×10^{-10}	8×10^{-8}	2.8×10^{-8}

used to determine values of β, assuming $\Phi_0=4.5$ (corresponding to the work-function for pure tungsten) and from these the values of Φ_0, a, and S_0 were calculated for the cases in which a film was present. The fact that S_0 re-. mained constant, while S, the apparent emitting area, varies with a and Φ_0, shows the validity of the theoretical relations deduced on the basis of wave mechanics.[207]

[206] Gosling, Phil. Mag. **1**, 609 (1926).

[207] In this connection attention may be drawn to the paper by P. W. Bridgman (Phys. Rev. **34**, 1411 (1929) in which is discussed the bearing of classical thermodynamics on emission in strong fields.

With regard to the effect of temperature on the auto-electronic emission, equation (72) gives the empirical relation suggested by Millikan and Eyring. However, W. V. Houston[208] has shown that on the basis of the same considerations as those used by Fowler and Nordheim, the emission as a function of both T and F differs from equation (76) only in the substitution for F^2 of the expression

$$\left\{F^2 + \frac{32\pi^4 mk^2}{3h^2} b_0 kT^2\right\}.$$

The temperature variation is thus quite small and decreases as F increases, so that for very large values of the latter, the emission is practically independent of T.[209]

XII. Thermal Emission of Positive Ions[210]

The earlier observations on positive ion emission from heated salts and metals have been discussed by O. W. Richardson in his book. The emission observed in these cases occurs, in general, at temperatures below those observed for electron emission; it decreases with time according to an exponential law and varies with the conditions of treatment of the emitter.

(A) Emission due to vapor of alkali metals.

As has been mentioned in connection with caesiated filaments, atoms of an alkali vapor striking an incandescent filament for which the work function, Φ_0, is greater than the ionizing potential, V_i, of the incident atoms are thereby ionized and if the filament is positively charged, a positive ion current is observed the magnitude of which (I_p) is practically independent of the filament temperature, above a certain value, and varies linearly with the pressure of the vapor in accordance with the equation[211]

$$I_p = pe/(2\pi mkT)^{1/2} \tag{79}$$

where m = mass of positive ion.

T = temperature of vapor corresponding to pressure p.

For caesium on tungsten, the lower limit of temperature is about 1100°K. As the temperature of the filament is decreased below this value, the alkali atoms are adsorbed more and more, thus decreasing Φ_0, until finally when $\Phi_0 \leqq V_i$ the positive ion emission is reduced to zero. The degree of ionization of caesium vapor at any temperature may be calculated by means of a modification of Saha's equation [212] which Langmuir and Kingdon have derived[213]

[208] W. V. Houston, Phys. Rev. **33,** 361 (1929).

[209] N. A. de Bruyne, Phys. Rev. **35,** 172 (1930) notes that he has failed to observe the effect of temperature stated by Millikan and Lauritsen.

[210] Although this topic has also been treated by Compton and Langmuir in their review, it has been considered advisable to summarize the observations in this field in the present paper.

[211] K. H. Kingdon, Phys. Rev. **23,** 774 (1924).

[212] Saha, Phil. Mag. **40,** 472 (1920); **46,** 534 (1923).

[213] Langmuir and Kingdon, Proc. Roy. Soc. **A107,** 61 (1925).

on the basis of the work of Fowler and Milne.[214] This equation has the form

$$\log_{10}\frac{n_e n_p}{n_a} = \frac{11,600V_i}{2.303T} + \frac{3}{2}\log_{10} T + 15.385 \qquad (80)$$

where n_e, n_p and n_a denote the numbers of electrons, positive ions and atoms respectively, per cm,^{3} V_i is the ionizing potential of the vapor in volts, and T is the absolute temperature.

The values of V_i for the different alkali and alkaline earth metals are as follows:

Li	5.37	Ca	6.09
Na	5.12	Sr	5.67
K	4.32	Ba	5.19
Rb	4.16	Cs	3.88

Hence, in presence of a tungsten filament ($\Phi_0 = 4.53$), only K, Rb and Cs may be expected to give positive ion currents. Equation (80) has been applied by Langmuir and Kingdon to calculate the degree of ionization of caesium atoms incident upon a heated tungsten filament. Under these conditions, the value of n_e is obtained from the electron emission, I_e, for pure tungsten by means of the relation

$$n_e = \frac{I_e}{e}\left(\frac{2\pi m_0}{kT}\right)^{1/2} = 4.034 \times 10^{13}\frac{I_e}{T^{1/2}} \qquad (81)$$

and this is then substituted in the value of $K = n_e n_p/n_a$ derived from (80) to obtain the ratio n_p/n_a. Thus for tungsten at 1200°, the calculated value of this ratio is 577, which means "that out of 578 caesium atoms incident on the filament, an average of only one will leave the surface as a neutral atom, while 577 leave as ions. This conclusion is in accord with the experiments."

On the other hand, if the caesium atoms strike a thoriated tungsten filament ($\Phi_0 = 2.69$) at $T = 1200°$K, $n_p/n_a = 8.9\text{x}10^{-5}$ which indicates negligible ionization. In their paper, Langmuir and Kingdon give values of n_e and n_p/n_a for a series of temperatures for caesium vapor in contact with tungsten and thoriated tungsten, and have also confirmed the validity of equation (80) by direct measurements of I_e and I_p.

Following up this work, T. J. Killian[215] has shown that the same theory is quantitatively applicable to potassium and rubidium in contact with tungsten and oxidized tungsten, and from the observations he has been able to derive vapor pressure curves for the metals, which are in agreement with those obtained by the more usual methods. It is also of interest to mention that a beautiful confirmation was obtained of the space charge equation for positive ions.

More recently E. Meyer[216] has reported on positive ion emission in potas-

[214] Fowler and Milne, Monthly Notices, Roy. Astr. Soc. **83**, 403 (1923). See discussion in review by Compton and Langmuir, p. 135.

[215] T. J. Killan, Phys. Rev. **27**, 578 (1926).

[216] E. Meyer, Inaug. Diss. Berlin (1930).

sium vapor as affected by variation in filament material. The results obtained for tungsten, molybdenum and tantalum were found to be in qualitative agreement with equation (80).

(B) Emission from catalysts.

The phenomenon of positive ion emission from certain catalysts is apparently best explained on the basis of the same theory as that used for the emission from tungsten and similar metals in caesium or rubidium vapor. C. H. Kunsman[217] observed that certain iron oxide crystals containing about 1 percent of some alkali metal or alkaline earth used in the Fixed Nitrogen Research Laboratory as a catalyst also provide a copious source of positive ions. Also it was found that the emission varies with the temperature in accordance with Richardson's equation.

H. A. Barton, G. P. Harnwell and C. H. Kunsman[218] have analyzed the ions emitted by means of the mass spectrograph and find that these ions are singly charged and have the same mass as that of the alkali or alkaline earth metal added to the iron oxide crystals. In the case of the alkaline earth metals a preliminary reduction of the mixture in hydrogen at a red heat was found to increase the emission. Platinum gave emission due to Na^+ and K^+ which diminished with time.

While the catalyst may be painted on platinum ribbon in the same manner as oxide coatings, a more convenient method devised in the Fixed Nitrogen Research Laboratory consists in the use of a thin-walled platinum tube with small perforations, in which the material is contained.

A potassium ion emitter prepared by a totally new method has been described by F. G. Cottrell, C. H. Kunsman and R. A. Nelson,[219] while A. K. Brewer[220] has discussed the photo-electric and thermionic properties of this source. The emitter consists of a platinum wire over which a thin layer of special potassium glass is fused. By electrolysis in molten potassium nitrate, a deposit of potassium is formed on this filament. The outer surface of the glass tube is then covered with platinum by cathodic sputtering, short distances being left uncoated at each end of the glass tube in order to provide insulation between the platinum deposit and the filament. By applying a potential between these, potassium can be electrolyzed to or from the external coating. The combination is used as an emitter for either electrons or positive ions. The emission of positive ions becomes detectable at 270° C. while that of electrons is detectable above 425° C. Also the electrode exhibits photo-electric sensitivity in the near ultraviolet with a threshold wavelength, of which the exact value depends upon the direction of current between the platinum coating and filament and on the temperature.

[217] C. H. Kunsman, Science **62**, 269 (1925); Proc. Nat. Acad. Sci. **12**, 659 (1926). Phys Rev. **27**, 249 (1926); Jr. Physical Chem. **30**, 189 (1926).

[218] H. A. Barton, G. P. Harnwell and C. H. Kunsman, Phys. Rev. **27**, 739 (1926).

[219] F. G. Cottrell, C. H. Kunsman and R. A. Nelson, J. Opt. Soc. Am. **20**, 152 (1930).

[220] A. K. Brewer, Phys. Rev. **35**, 1360 (1930).

(C) Emission of positive ions of metals.

Another type of positive ion emission observed by H. B. Wahlin[221] and L. P. Smith,[222] is that in which the ions are charged atoms of the electrode material itself. This has been demonstrated by Wahlin by measurements with the mass spectrograph in the case of chromium, molybdenum, tungsten, rhodium, ruthenium, tantalum and columbium. All these metals give alkali metal ions when first heated, but this disappears after a few minutes. In the case of iron, nickel, copper, zirconium, platinum and palladium the emission of alkali ions persists for a much longer time. The emission of positive ions of the metals becomes measurable when the temperature is raised to a point at which appreciable evaporation occurs, and L. P. Smith has investigated more extensively this ion emission from tungsten and molybdenum.

The variation in emission with temperature is observed to follow the equation for electron emission. From the plot of log (I/T^2) vs. $1/T$ the work function for positive ion emission is found to be 6.55 volts for tungsten and 6.09 volts for molybdenum, whereas the values calculated from a thermodynamical energy cycle are 11.08 and 10.51 volts respectively. Again the integration constant A is actually found to be about 10^6 times smaller than that calculated by a method similar to that used by Dushman and Bridgman in calculating the theoretical value of A_0 for electron emission. Regarding the cause of these differences, no satisfactory explanation is apparent at present. Consequently it must be concluded "that the mechanism of emission of electricity from hot metals is not clearly understood, especially as regards the emission of positive ions, for, unlike the evaporation of electrons and neutral atoms, the evaporation of ions is not strictly represented by an equation based upon thermodynamical arguments under equilibrium conditions."

[221] H. B. Wahlin, Phys. Rev. **34**, 164 (1929); Langmuir and Compton, loc. cit.

[222] L. P. Smith, Phys. Rev. **34**, 1496 (1929); **35**, 381 (1930). This gives references to previous literature on this subject.

MAGNETISM IN DISCONTINUOUS MEDIA

By L. W. McKeehan
Yale University

Introduction

MAGNETS have always interested the inquisitive, perhaps the more so because the peculiar state of things in the neighborhood of magnets is not directly appreciable by any of our sense organs. Since the discovery by Oersted of a connection between magnets and electric currents, and the immediately succeeding researches, by Ampere and by Faraday, which defined this connection and made it possible to use electrical apparatus in the study of magnetism, the problems magnetism presents have been attacked again and again, and the results of experiment have been discussed with great gusto by many writers of books and monographs. The present situation is such that the veriest tyro in physics can easily collect new, and possibly significant, data which the most learned can hardly explain in terms of simple concepts. Under these conditions, which have long obtained, it is not surprizing that experiment has here outstripped theory farther than in any other part of the physicist's universe. It is the aim of this review to call to the attention of those concerned the adequacies and deficiencies of present-day attempts at explanation, and to suggest some promising lines for further advances.

The title of this review has been chosen to emphasize the fact that we are now thoroughly committed to the type of theory which denies the existence of those continuous jelly-like media with which alone the Victorian school of theoretical physicists, and its predecessors, would have anything to do. There are not, as far as we need now to suppose, any media which are absolutely discontinuous in a mathematical sense. All real media, on the other hand, are physically discontinuous in the sense that they naturally subdivide themselves into regions where we can at least imagine ways of experimenting that would disclose relatively slow changes of condition from point to point, or from instant to instant, and regions where changes are so abrupt that no sufficiently delicate probe can be inserted into them, even in imagination, to analyse their peculiarities. The boundaries, necessarily rather hazy, between regions of these two sorts are the discontinuities referred to in our title. We have not, in choosing it, denied ourselves the right to consider any fact of experience, but in the interest of brevity we will take care to pass lightly over those phenomena for which the older continuous fluid theories are still adequate, in order to spend more time with experiments and explanations which stress the atomistic and catastrophic features of our subject.

The Magnetic Field

Magnetic theory made its first great advance when it introduced the notion of the magnetic field. The magnetic field intensity, or magnetic force,

$\boldsymbol{H}$, is a vector point function which represents in every accessible region the resultant magnetic effect of all the magnets and electric currents with which we may for the moment be concerned. In attempting to define this field intensity in terms of electric charge and its motion—the only undefinable fundamentals which we allow ourselves in this branch of physics—we find it necessary to introduce an electric field intensity, $\boldsymbol{E}$, which is also a vector point function.

The familiar equations stating all that we know about $\boldsymbol{H}$ are:

$$\nabla \cdot \boldsymbol{H} = 0,$$

$$\dot{\boldsymbol{H}} = -c(\nabla \times \boldsymbol{E}),$$

$$\nabla \times \boldsymbol{H} = \frac{1}{c}(\dot{\boldsymbol{E}} + 4\pi\rho\boldsymbol{v})$$

The additional equation necessary for defining $\boldsymbol{E}$ is:

$$\nabla \cdot \boldsymbol{E} = 4\pi\rho.$$

In these equations ρ is the volume density of positive electric charge and v is its velocity at the point considered; c is the velocity of light. The quantity H will be in e.m.u. (gauss) if E is in e.s.u., (which puts ρ in e.s.u. cm^{-3}), v and c in $\text{cm} \cdot \text{sec.}^{-1}$ The somewhat unusual order of the equations has been chosen to emphasize our pre-occupation with $\boldsymbol{H}$.

Not many years ago those parts of space which were, as the phrase went, occupied by ponderable matter, were regarded as inaccessible to detailed study, and the electromagnetic field equations were revised in such a way that the gross phenomena of magnetism might be expressed by equations almost as simple as those written down in the last paragraph.

This was a bold and clever move, but in the sequel it has made a great deal of trouble, particularly between physicists and electrical engineers, and since all this trouble is due to the fallacious assumption that a magnetic medium is a continuous jelly with any properties we choose to foist upon it, we will do well to consider briefly what went wrong.

Faraday's "tube of force" was undoubtedly the root of all evil in this connection. To preserve these tubes unaltered in number as they were followed through various pieces of matter about the interior of which nothing much was known, it was found necessary to invent for them a vector point function, $\boldsymbol{B}$, called the magnetic induction, which could not be distinguished from $\boldsymbol{H}$ in what was fondly termed "empty space," but which differed markedly from $\boldsymbol{H}$ in "ponderable matter." To make matters worse, it soon appeared desirable to invent two more vector point functions for ponderable matter only, and one of these, in all innocence, was called the magnetic field intensity and given the symbol $\boldsymbol{H}$ in spite of the fact that it did not satisfy the equations—given above—by which alone $\boldsymbol{H}$ had been defined.[1]

[1] The order in time of these *faux pas* has been altered to make their cumulative absurdity more apparent.

The other new vector point function, the intensity of magnetization, I, did not usurp a place already occupied.

In order to reduce by a little the resultant confusion we will here distinguish this interloping H as H', and call it the *fictitious* magnetic field intensity. It represents, at a point inside a "magnetized body" the magnetic field intensity which would be found there if the body were replaced by a vacuum without changing the state of external magnets or currents in any way. In order at the same time to preserve the values of H, due in part to the thus annihilated body, at external points, it is necessary to distribute over its vacated surface a proper distribution of (fictitious) free magnetic poles. The magnetic field intensity which would then arise at internal points from the presence of these (fictitious) poles is $4\pi I$. In using these definitions for H' and I it should always be remembered that the "body" to be considered is only so much of the real piece of matter as is strictly homogeneous in its magnetic behavior. In many cases the fictitious magnetic poles would have to be spread throughout the volume as well as over the bounding surface of the piece. With all these precautions then, we may use when we like the following equations:

$$B = H' + 4\pi I,$$
$$B = \mu H',$$
$$I = \kappa H'.$$

Here μ and κ are, respectively, the magnetic permeability and the magnetic susceptibility. Both μ and κ may be complex numbers. It is of no interest to us here whether μ and κ are pure numbers or not, nor, whether B and H need separate names for the units in terms of which they are expressed.[2] We can see from these equations, and it is worth emphasizing the point, that B is a space-average of H throughout a small region which is nevertheless large enough to include representative samples of all the various values of H which the most refined analysis could discover.

We must be careful to note that though we may never probe deeply enough into atomic structure to interest ourselves in the intense and strongly curved fields of H which we suppose to exist inside atoms and nuclei, we may still suppose that H has a value at every point in space and use our electromagnetic field equations in more accessible parts of magnetic media without any fear of self-contradiction.

Elementary Magnets

The second great impetus to magnetic theory was afforded to it by Ampere's recognition of the fact that the magnetic effects of permanent and in-

[2] Just as this goes to press the International Electrotechnical Commission is reported—Nature **126**, 252 (1930)—to have adopted "gauss" as the name for the unit of flux density, *i.e.* of **B** and **H** as used herein, and "oersted" as the name for the unit of magnetic field intensity [inside ferromagnetic media], *i.e.* of **H'** as used herein. The second of these names has not been current in this sense, and unless the confusion between **H** and **H'** gets to be generally understood, it is hardly likely to become so.

duced magnets are attributable to the presence within them of permanently circulating electric currents. Weber added the notion that non-decaying eddy currents can account for diamagnetic behavior, and suggested that both sorts of currents may circulate within the molecule, thus constituting truly elementary magnets.[3] The assumed structure of these Weber elements has changed from time to time as fashion has dictated. They began as resistanceless grooves in which continuous streams of electricity were supposed to flow. At a later period they became spherical or ellipsoidal clouds of negative electricity in rotation about their axes in a more rarefied cloud of positive electricity.[4] Still later they metamorphosed into electrons in orbital motion. At present they are usually conceived to be spinning electrons, the centers of which are found from instant to instant to lie at any distance—but with certain definite distances more highly probable than others—from an atomic nucleus also capable of spin. Concerning the spatial distribution of the electric charge which spins, theories are now rather non-committal, ignoring the possibility of spatial definiteness within sub-atomic regions. This seems rather unkind when we remember that the anomaly in the gyromagnetic ratio—the ratio of the angular momentum to the magnetic moment of the moving electricity—was at one time taken to be strong evidence in favor of the hypothesis that elementary magnets were spinning electrons.

The least magnetic moment of a modernized Weber element is agreed to be about $9.2 \cdot (10)^{-21}$ e.m.u.[5] (The magnetic moment of an atomic nucleus is supposed to be much less than this.[6]) This magnetic moment is also a fundamental unit; a Weber element may have a magnetic moment equal to any integral multiple of these so-called Bohr magnetons. The designation "Bohr" is necessary because of the prior existence of an empirical unit, the "Weiss" magneton, the present agreed value of this being $1.857 \cdot (10)^{-21}$ erg $\cdot$ gauss^{-1} $\cdot$ mol^{-1}. Weber elements with more than a few Bohr magnetons are as yet unknown.

The least magnetic moment associated with matter is that exhibited by an atom, ion or molecule, which may contain from one to several hundred Weber elements with from one to six Bohr magnetons each. It might be supposed that the resultant magnetic moment of structures as complicated as some of these would be a very indefinite quantity with a fair probability of high values in terms of the unit. It appears, however, that the stable configurations are of such a sort that the net magnetic moment is always a small

[3] W. Peddie gives in his Molecular Magnetism, Longmans Green and Co. 1929, pp. 17–22, a clear account of Weber's contributions to magnetic theory. What Weber called a molecule would now more usually be identified as an atom, especially in crystals of the chemical elements.

[4] R. Gans, Ann. d. Physik [4] **49**, 149–178; **50**, 163–198 (1916).

[5] R. T. Birge—Phys. Rev. Suppl. **1**, 1–73 (1929)—gives $(9.174_{70} \pm 0.013) \cdot (10)^{-21}$ erg $\cdot$ gauss^{-1} for the value derived from the best spectroscopic data, $(9.216_{38} \pm 0.016) \cdot (10)^{-21}$ erg $\cdot$ gauss^{-1} for the value derived from experiments on the deflection of cathode rays.

[6] E. Fermi, Zeits. f. Physik **60**, 320–333 (1930); J. Dorfmann, Zeits. f. Physik **62**, 90–94 (1930); S. Goudsmit, L. A. Young, Phys. Rev. [2] **35**, 1418–1419 (1930); H. S. W. Massey, Proc. Roy. Soc. [A] **127**, 666–670 (1930).

integral number of Bohr magnetons. In very many cases the net magnetic moment is zero, in no case yet studied is it more than seven Bohr magnetons.[7]

The behavior of isolated atoms, ions and molecules is thus about as simple to handle as that of isolated Weber elements. In what follows we will therefore be justified in discussing such units of matter, when independent of each other's influences, as if they were small rigid magnets of moment $\boldsymbol{M}$, the scalar part of which, M, is a constant.

The simplest possible ion, an isolated electron, ought, unless the phenomenon of spin can in some way be denied to it, to have a magnetic moment of one Bohr magneton, and this may modify its motion in a predictable way under assigned conditions. Several experiments to settle this point have been imagined,[8] and the results of experiments on the scattering of beta-particles[9] have been taken to throw some light on the subject. No conclusion can as yet be drawn.

Wherever the units of matter have any freedom of rotation, that is under all known conditions, we must expect diamagnetism to exist. If any of the units have values of M different from zero paramagnetism must exist. If any of these units are so constrained that $\boldsymbol{M}$ can maintain its direction against disturbing agents, ferromagnetism is present. It will be noticed that diamagnetic, paramagnetic, and ferromagnetic are not exclusive categories. It is none the less convenient to consider such properties separately and in the order just given because the conditions become increasingly complex as we thus proceed from the general to the particular. Naturally enough we find more and more convincing evidence for the atomicity of magnetism as we proceed in this direction.

A cross-classification, which is useful in separating the experimental and theoretical work into more homogeneous groups, is that afforded by the physical state of the system considered. We will follow this plan, first dealing with isolated atomic magnets, then with fluid media and finally with solids, that is to say, with crystals. In the treatment of crystals we will take up, not quite in chronological order, the more important theories so far proposed, prefixing the name of its author to the section in which each theory is treated.

Isolated Atomic Magnets

If an atom or molecule be shot through a non-uniform magnetic field its direction of motion will change in a manner and to an amount which depends upon its magnetic moment, $\boldsymbol{M}$. This is an abbreviated description of the famous Stern-Gerlach experiment, which may be said to have demonstrated the existence of atomic magnets. In the hands of its originators and their

[7] J. Becquerel and W. J. de Haas—J. de phys. [7] **1** 7S–9S (1930)—assign this magnetic moment to the ions of gadolinium in the mineral xenotime.

[8] L. Brillouin, Comptes rendus, **184,** 82–84 (1927); Proc. Nat. Acad. Sci. **14,** 755–763 (1928); V. M. Albers, T. E. Phipps, Phys. Rev. [2] **31,** 149 (1928); T. Sexl, Zeits. f. Physik **48,** 729–730 (1928); F. Wolf, Zeits. f. Physik **52,** 314–317 (1928); F. E. Myers, R. T. Cox, Phys. Rev. [2] **34,** 1067–1068 (1929); F. Knauer, Zeits. f. Physik **59,** 807–811 (1930).

[9] M. C. Henderson, Phil. Mag. [7] **8,** 847–857 (1929).

followers it has already supplied much evidence confirming the spectroscopist's assignment of electrons to the various groups within the atomic structure.[10] Changing the type of non-uniform field, and the way of entering or leaving its boundaries, should yield additional information.[11]

The Stern-Gerlach experiment is a brilliant example of success where success was hardly to be expected. It is easy to prove that a magnet fired from field-free space into a magnetic field of any sort can never align itself with respect to the field, but must continually precess so as to maintain invariant that component, M_p, of its original magnetic moment, $\boldsymbol{M}$, which lies parallel to the field. The proof being due to Larmor such precession is referred to under his name. All that classical theory can predict under the circumstances is a symmetrical broadening of the beam in the direction of the gradient of $\boldsymbol{H}$, here roughly parallel to $\boldsymbol{H}$ in the strongest part of the field. The actual effect is a splitting of the beam into several beams in each of which M_p is an integral number of Bohr magnetons, including $+M$ and $-M$ among its values.

The logical discontinuity here presented has been encountered in other places, but seldom so patently. To escape from the dilemma it is necessary to jettison classical electro-dynamics.

Consider a Weber element entering a magnetic field and beginning to precess. The system formed by the field and the magnet is characterized by two frequencies, one being that of the rotation which fixes M, the other being that of the Larmor precession. Since the frequency of precession is low in the weak magnetic field first traversed, the corresponding energy states are, on quantum theory, close together and the probability of transition from one state to another is high. The magnetic energy of each state is $W_i = \boldsymbol{M}_i \cdot \boldsymbol{H}$, where the vectors $\boldsymbol{M}_i$ differ only in direction. We may also consider the fact that the states are ill-defined—the quantization weak—because the magnet does not stay long enough in any one part of the field for the precession to complete many cycles at substantially constant frequency. Under these conditions the system radiates energy in small amounts so as continually to adjust its energy to one of the states in which the value of M_p is integral in Bohr magnetons. In the stronger parts of the field the quantization is stronger and the stability of any permissible state is correspondingly greater. Here the resolution of the beam into its components of different M_p has time to occur. Finally, in leaving the field, the quantization again weakens and the individual beams become more diffuse. The greater the number of permissible states—the greater the value M—the less is the chance for clear resolution. Thus Gerlach[10] now states that the results for iron atoms do not, as at first supposed, indicate $M = 0$, but suggest that $M > 6$. (In the case of cobalt a value $M_p = 6$ can just be established.) The experimental resolving

[10] W. Gerlach—J. de phys. [6] **10**, 273–282 (1929)—has recently summarized the situation and given some further information regarding the first experimental results.

[11] I. I. Rabi, Zeits. f. Physik **54**, 190–197 (1929); D. E. Olshevsky—Phys. Rev. [2] **35**, 659 (1930)—is preparing a critical summary of all possible methods.

power is always inferior to the theoretical limit because all the atoms entering the field do not, as has been tacitly assumed, have equal velocities.

The behavior of a diamagnetic atom or molecule under the conditions of the Stern-Gerlach experiment is quite different from that of the paramagnetic unit so far considered. Since the magnetic moment of the diamagnet is, except for a small correction term,[12] directly proportional to H and antiparallel to $\boldsymbol{H}$, all the units of one kind should follow the same path and no resolution into discrete beams can be expected. It is still possible to imagine that a gas may permanently contain molecules with different moments of inertia or with different distributions of magnetically balanced Weber elements.[13] If so, the experiment should give pertinent data regarding these differences.

Fluids

The molecules of a gas under ordinary conditions of experiment may be regarded as magnetically independent. The magnetization, I, of a gas-filled space is therefore dependent upon the distribution of the unitary magnets among the same permissible energy states as those of interest in the Stern-Gerlach experiment and its derivatives. Explicitly,

$$I = \frac{1}{V}\sum_{V} M_p,$$

where V is a small volume containing many unitary magnets. This additive property of the units has a very important consequence. It permits us to study the magnetic moment of many molecules and ions in solution, by making proper correction for the magnetic characteristics of the solvent.

Even in complex molecules, and particularly in organic compounds, it is found that the atoms are often sufficiently free to behave independently as regards magnetic matters, so that Pascal[14] and others have been able to assign to each of several important constituent atoms its proper contribution to the magnetic result and to show that the mode of combination has a characteristic effect upon this contribution.[15] In heteropolar compounds, and in what are called coordination compounds, the magnetic contribution of a given chemical element may vary not only in magnitude but even in sign.

In a diamagnetic fluid the parallelism of the induced magnetic moments of the units makes it difficult to find out anything about the individuals from a study of the magnetic susceptibility of the medium they constitute. For a time it seemed that an anomalous variation of susceptibility with pres-

[12] J. H. Van Vleck, A. Frank, Proc. Nat. Acad. Sci. **15**, 539–544 (1929); F. Bitter, Phys. Zeits. **30**, 497–504 (1929); K. F. Niessen, Phys. Rev. [2] **34**, 253–278 (1929).

[13] E. C. Stoner—Proc. Leeds Phil. Lit. Soc. [Sci. Sect.] **1**, 484–490 (1929)—discusses the effect of spatial distribution of electric charge upon diamagnetic susceptibility.

[14] P. Pascal, Ann. de chim. et de phys. [8] **16**, 359–402, 520–574 (1909); **19**, 5–70 (1910); **25**, 289–377 (1912); **29**, 218–243 (1913).

[15] The comparison of electronic isomers is particularly interesting: S. Bhatnagar, R. N. Mathur, R. S. Mal, Phil. Mag. [7] **10**, 101–109 (1930).

sure—observed by Glaser—would furnish such detailed information. Other experimenters[16] attribute the anomaly to accident, but the discoverer still[17] thinks it significant. The wave mechanical theory of diamagnetism[12] offers some hope to the experimenter by predicting the presence of a paramagnetic term in the susceptibility if the electron distribution is not of spherical symmetry, but this term must generally be small even in comparison with the difficult measurable diamagnetic term.

In paramagnetic solutions the observed magnetic moment per dissolved atom may give very valuable indications of its electronic structure. Weiss[18] and Cabrera[19] have furnished most of the data, and attempt to show that the results are compatible with the existence of a unit of magnetic moment, the so-called Weiss magneton, which has been mentioned above. There are several puzzling cases, however, in which it has been necessary to suppose that ions of one sort may, under one set of conditions, have either of two different magnetic moments, each an integral multiple of the preferred unit. By this means non-integral multiples can be explained away. In the interests of fair play it should be remarked that the same artifice is even more often necessary in explaining experimental results in terms of the five times larger Bohr magneton. In this case, however, the argument for a variety of integral contributions, including zero, is part of the original theory and not an *ad hoc* addition. Much ingenuity has been displayed in bringing the quantum theory into agreement with the curious succession of M-values found for ions having from 18 to 28 extra-nuclear electrons—the iron family—and those having from 54 to 68—the rare-earth group. It will be sufficient here to remark that the agreement now appears remarkably exact.[20]

The value of M is obtained by what amounts to an extrapolation to infinite field intensity or to zero absolute temperature. The law of variation of paramagnetic susceptibility with H and T lies therefore at the bottom of all the experimental and theoretical agreements just mentioned. This fundamental relation was derived by Langevin on the assumption that the molecules of a gas were small permanent magnets without transverse dimensions, so that the principal moment of inertia about the magnetic axis of each was zero. The treatment was classical so that it was assumed that all magnetic

[16] G. W. Hammar, Proc. Nat. Acad. Sci. **12**, 594–596, 597–598 (1926); V. I. Vaidyanathan, Indian J. Physics **1**, 183–198 (1926); **2**, 135–150 (1928); Phil. Mag. [7] **5**, 380–386 (1928); H. Buchner, Ann. d. Physik [5] **1**, 40–48 (1929); F. Bitter, Phys. Rev. [2] **35**, 1572–1575 (1930).

[17] A. Glaser, Ann. d. Physik [5] **2**, 233–248; **3**, 1119–1135 (1929); **4**, 82–90 (1930).

[18] P. Weiss, Arch. des sci. [4] **31**, 401–438 (1911); **35**, 406–408 (1913); [5] **2**, 417–419 (1920); J. de phys. [6] **5**, 129–152 (1924); Helv. Phys. Acta **2**, 283–284 (1929); Trans. Amer. Electrochem. Soc. **55**, 75–78 (1929); J. de phys. [7] **1**, 185–193 (1930); F. Birch, J. de phys. [6] **9**, 137–141 (1928).

[19] B. Cabrera, J. de phys. [6] **3**, 443–460 (1922); Anal. fiis. quim. **21**, 505–526 (1923); **23**, 151–157 (1925); J. de phys. [6] **6**, 241–258, 273–286 (1925); **8**, 257–275 (1927); Atti congr. internat. dei fisici, Como 1927, **1**, 95–115 (1928); Anal. fiis. quim. **26**, 50–66 (1928).

[20] D. M. Bose, Atti congr. internat. dei fisici, Como 1927, **1**, 119–125 (1928); Phil. Mag. [7] **5**, 1048–1067 (1928); D. M. Bose, H. G. Bhar, Zeits. f. Physik **48**, 716–721 (1928); E. C. Stoner, Phil. Mag. [7] **8**, 250–266 (1929); A. Sommerfeld, Wien. Ber. **139** [2a] 11–17 (1930).

states—all relative orientations of a magnet and the applied magnetic field —were permissible. The temperature dependence of paramagnetic susceptibility arises on this theory from the increasing number of collisions with increasing temperature at constant volume.

It is apparent at a glance that the magnets of Langevin's theory differ in one important respect from the magnetically unbalanced assemblages of Weber elements to which we have committed ourselves in the present discussion. The magnets in our picture must possess angular momenta, those in Langevin's picture must not, for if they did the applied magnetic field—on classical dynamics—could never align them. This contradiction, which we found it necessary to circumvent in the Stern-Gerlach experiment has most recently been emphasized by Honda,[21] but has not escaped previous remark by others.[22] The fact that Langevin's theory works so well for extrapolation is another of those pleasant surprizes which make experimental physics so exciting.

The substitutes for Langevin's theory which are acceptable in quantum mechanics differ from it principally under extreme conditions.[23] The new expression for relative intensity of magnetization in terms of the saturation value,

$$\frac{I}{I_s} = \tanh \frac{MH}{kT},$$

can hardly be distinguished for wide ranges of H and T from Langevin's expression

$$\frac{I}{I_s} = \coth \frac{MH}{kT} - \frac{kT}{MH}.$$

Schidlof[24] has recently discussed the range of validity of the quantum-theoretical result and makes it clear that pushing the equation too close to $T=0$ is not justifiable. The practice of extrapolating Langevin's equation into the same region is still more unwise.

Meso-Phases

A few solids do not melt to form clear liquids but have a turbid phase which is stable over a definite range in temperature. These meso-phases are classified as nematic—stringy—or smectic—smeary—depending upon the structures they exhibit when subjected to the action of magnetic fields, adjacent surfaces, or other orientating agent. The control of their structure by magnetic fields is formally attributable to their anisotropy of magnetic susceptibility. The axis of the strings in nematic fluids and the normal to

[21] K. Honda, Zeits. f. Physik **63**, 141–148 (1930). In an earlier paper—Sci. Rep. Tohoku Imp. Univ. [I] **17**, 997–1009 (1928); Zeits. f. Physik **47**, 691–701 (1928)—he avoided this difficulty, in ferromagnetic media only, by inventing a nuclear magnet without angular momentum.

[22] *e.g.* by J. Kroo, Ann. d. Physik [4] **42**, 1354–1396 (1913).

[23] W. Lenz, Phys. Zeits. **21**, 613–615 (1920); W. Schütz, Zeits. f. Physik **54**, 731–735 (1929).

[24] A. Schidlof, Helv. Phys. Acta **1**, 578–600, 601–612 (1928).

the layers in smectic fluids are unique axes of magnetization. An attempt has been made[25] to compute the value of M for a magnetic unit which would be orientated in the observed manner in spite of heat motions, but the assumption that the only controlling force is magnetic in origin is probably not allowable.[26] Besides more recent studies on the magnetic properties of these interesting substances themselves,[27] an attempt has been made to modify the properties of the crystals which freeze out from them by applying a strong magnetic field during their formation.[28] The result of this test was negative in the case of a smectic substance, positive in the case of a nematic substance.

Diamagnetic Crystals[29]

The first attempt to construct a theory for the magnetic behavior of crystals was made by William Thomson (later Lord Kelvin) in 1850.[30] His theory took no account of the atomicity of magnetism, and is of no interest to us here except for the fact that it predicts that crystals of the cubic system are magnetically isotropic. Certain cubic crystals have been reported to be anisotropic by Forrest[31] but contrary results for one of the simplest of the materials he tested—copper—have more recently been obtained by Montgomery[32] in this laboratory.

In non-cubic diamagnetic crystals the anisotropic susceptibility makes it necessary to suppose that the non-decaying eddy currents are constrained to move in particular planes with respect to the crystal axes. This means, in turn, that the atom, or larger structure, through which these currents circulate is prevented by its neighbors in the crystal from rotating freely about a random axis.

Recent experiments by Raman[33] and his colleagues[34] make it appear probable that in bismuth, antimony and graphite the abnormally high diamagnetic susceptibility along the three-fold or six-fold axis is due to circuital currents which encircle not one but many atoms. Subdivision of the crystals,

[25] W. Kast, Ann. d. Physik [4] **83,** 391–417 (1927).

[26] H. Zocher, Phys. Zeits. **28,** 790–796 (1927); M. Jezewski, Zeits. f. Physik **52,** 268–274, 878–881 (1929).

[27] V. Freederiksz, V. Zolina, Trans. Amer. Electrochem. Soc. **55,** 85–96 (1929); A. v. Wyk, Ann. d. Physik [5] **3,** 879–893 (1929); H. Zocher, V. Birstein, Zeits. f. phys. Chem. [A] **142,** 186–194 (1929).

[28] G. Foëx, J. de phys. [6] **10,** 421–438 (1929).

[29] Non-crystalline solids may be regarded as hyper-viscous fluids and demand no special treatment here. Disorderly aggregates of crystals are considered later. Unless otherwise stated the crystals considered by theory are ideally perfect.

[30] Rep. Brit. Assoc. [II], 23 (1850); Phil. Mag. [4] **1,** 177–186 (1851).

[31] J. Forrest, Trans. Roy. Soc. Edinb. **54,** 601–701 (1926); Phil. Mag. [7] **3,** 464–476 (1927).

[32] C. G. Montgomery, Phys. Rev. [2] **36,** 498–505 (1930).

[33] C. V. Raman, Nature **124,** 412 (1929); Proc. Phys. Soc. London **42,** 309–320 (1930); *cf.* P. Ehrenfest, Physica **5,** 388–391 (1925); Zeits. f. Physik **58,** 719–721 (1929).

[34] V. I. Vaidyanathan, Nature **124,** 762 (1929); **125,** 672 (1930); S. Paramasivan, Indian J. Phys. **4,** 139–145 (1929).

or heating, makes the abnormality less. The effect of heating is signfiicant because, unless inter-atomic supraconduction of the postulated sort is broken up by heat motions temperature changes should not affect the atomic susceptibility.

The alkali metals, for which theory can make some prediction of atomic susceptibility, have been studied with particular care by Lane[35] and by McLennan and his co-workers,[36] who have also made measurements on zinc and cadmium.[37] Bismuth in massive crystals has been reexamined by Fock.[38] Several investigators[39] have reported on crystals of organic compounds or on other classes of compounds where related structures may be compared. Bose[40] has found a simple salt of nickel which is diamagnetic. The effect of alloying has been examined by Davies and Keeping.[41] All these data will find their place in the final theory for the dependence of diamagnetism on crystal structure. Other effects which are as yet outside any theory are changes in the thermoelectric behavior of bismuth when crystallized in a strong magnetic field,[42] the effect of cold work upon the susceptibility of copper and silver,[44] and the changes in the magnetic behavior of solid solutions of copper and gold with heat treatment which favors a regular distribution of each constituent.[45]

Paramagnetic Crystals

In many paramagnetic crystals the susceptibility, independent of direction, may be interpreted in terms of Langevin's theory, or its quantum substitute, as well as if the medium were a fluid containing the same average concentration of unitary magnets. Most of the experimental work has been done at very low temperatures, where the dependence of susceptibility upon

[35] C. T. Lane, E. S. Bieler, Trans. Roy. Soc. Canada [3] **22** [III], 117–123 (1928); C. T. Lane, Diss. McGill (1929); Phil. Mag. [7] **8,** 354–362 (1929); Phys. Rev. [2] **35,** 977–981 (1930); for the theory see W. Pauli, Zeits. f. Physik **41,** 81–102 (1927); L. Rosenfeld, Naturwiss. **17,** 49–50 (1929).

[36] J. C. McLennan, R. Ruedy, E. Cohen, Proc. Roy. Soc. [A] **116,** 468–483 (1927); J. C. McLennan, E. Cohen, Trans. Roy. Soc. Canada [3] **23,** [III], 159–168 (1929).

[37] J. C. McLennan, R. Ruedy, E. Cohen, Proc. Roy. Soc. [A] **121,** 9–21 (1928).

[38] A. B. Focke, Phys. Rev. [2] **36,** 319–325 (1930).

[39] S. Bhagavantam, Proc. Roy. Soc. [A] **124,** 545–554; **126,** 143–154 (1929); Indian J. Phys. **4,** 1–14 (1929); S. S. Bhatnagar, S. L. Luther, J. Indian Chem. Soc. **6,** 303–307 (1929); S. S. Bhatnagar, R. N. Mathur, Phil. Mag. [7] **8,** 1041–1055 (1929); R. Hocart, Comptes rendus, **188,** 1151–1153 (1929); C. V. Raman, Nature **123,** 605, 945 (1929); F. W. Gray, J. Farquharson, Phil. Mag. [7] **10,** 191–216 (1930).

[40] D. M. Bose, Nature **125,** 708 (1930).

[41] W. G. Davies, E. S. Keeping, Phil. Mag. [7] **7,** 145–153 (1929).

[42] A. Goetz, R. C. Hergenrother, A. B. Focke, Phys. Rev. [2] **34,** 546–547 (1929); A. Goetz, M. Hassler, Phys. Rev. [2] **34,** 549 (1929); A. Goetz, Phys. Rev. [2] **35,** 193–207 (1930). In other crystals so far tested this procedure is without effect: L. A. Welo, Phys. Rev. [2] **34,** 296–299 (1929).

[43] P. L. Kapitza, Nature **124,** 53 (1929).

[44] F. Bitter, Phys. Rev. [2] **36,** 978–983 (1930).

[45] H. J. Seemann, E. Vogt. Ann. d. Physik [5] **2,** 976–990 (1929).

magnetic field intensity becomes important. The Cryogenic Laboratory of the University of Leiden has led in these investigations.

Becquerel and de Haas[46] have recently discovered that the Faraday effect —the rotation of the plane of polarization of light transmitted in the direction of an applied magnetic field—may be used as a measure of the magnetization in transparent crystals. This permits quantitative studies with specimens too small for examination by electrical methods.

Langevin's theory predicts that in low magnetic fields paramagnetic susceptibilities should vary inversely as the absolute temperature. This agreed with the previous experiments of Pierre Curie upon what he regarded as typical paramagnetic substances. As data accumulated it became apparent that many paramagnetic crystals and crystal aggregates show a more complicated dependence upon temperature.[47] Weiss pointed out that the form of Curie's Law could, at least in many cases, be preserved by measuring temperature not from absolute zero but from an origin chosen to give the best fit with the data. In the case of media which, at low temperatures, are ferromagnetic, this arbitrary origin as derived from high-temperature paramagnetism agrees fairly well with the temperature (Curie point, T_c) at which ferromagnetism finally disappears on heating. It is therefore customary to describe the temperature at which the value of the paramagnetic susceptibility, extrapolated from higher temperatures, would become infinite, as the paramagnetic Curie point.[48]

Unfortunately for the simplicity of the Curie-Weiss law, the paramagnetic Curie point is often below absolute zero, and even when it lies high enough to be accessible it is found that the law of variation of susceptibility changes as the critical temperature is approached, most substances remaining merely paramagnetic however much they may be cooled.[49] Weiss interprets such data as indicating allotropic modifications at definite temperatures, the low-temperature modification possessing a paramagnetic Curie point different from that of the high-temperature form—lower if ferromagnetism is successfully evaded. Others prefer to regard these results as proving that the simple inverse proportionality of atomic susceptibility and "corrected" temperature is an approximation at best, which diverges more and more

[46] J. Becquerel, Proc. Amst. **32,** 523–535, 749–758 (1929); Zeits. f. Physik **58,** 205–216 (1929); J. de phys. [6] **10,** 313–320 (1929): J. Becquerel, W. J. de Haas, Proc. Amst. **32,** 536–549, 578–590, 590–596, 1199–1205 (1929); Zeits. f. Physik **57,** 11–29 (1929); J. de phys. [6] **10,** 283–293 (1929); [7] **1,** 7S–9S (1930); H. A. Kramers Proc. Amst. **32,** 1176–1189 (1929); H. A. Kramers, J. Becquerel, Proc. Amst. **32,** 1190–1198 (1929); J. Becquerel, W. J. de Haas, H. A. Kramers, Proc. Amst. **32,** 1206–1211 (1929); *cf.* M. Ramanadham, Indian J. Phys. **4,** 15–38, 109–129 (1929); L. Rosenfeld, Zeits. f. Physik **57,** 835–854 (1929).

[47] In some cases the paramagnetic susceptibility does not vary with temperature to an appreciable extent: P. Weiss, P. Collet, Comptes rendus **178,** 2146–2149 (1924); P. Collet, Comptes rendus **183,** 1031–1032 (1926); N. Perrakis, J. de phys. [6] **8,** 473–480 (1928); Trans. Amer. Electrochem. Soc. **55,** 79–83 (1929); A. Serres, Comptes rendus **188,** 1239–1241 (1929), S. Holgersson, A. Serres, Comptes rendus **191,** 35–37 (1930).

[48] The discrepancies between ferromagnetic and paramagnetic Curie points, T_f and T_p, have been discussed by R. Forrer, J. de phys. [7] **1,** 49–64 (1930).

[49] H. R. Woltjer, E. C. Wiersma, Proc. Amst. **32,** 735–738 (1929).

from the facts as the system becomes more and more degenerate—less and less "classical"—with decreasing temperature. The most elaborate analysis of this type is that offered by Schidlof.[24] His equations predict that at very low temperatures the paramagnetic susceptibility should increase very slowly, by only one-third of its value—here finite—at the pseudo Curie point—here always positive—in passing from that temperature to absolute zero.

In non-cubic crystals the susceptibility in different crystallographic directions is different, and varies differently with temperature, so that one crystal may have several paramagnetic Curie points and several Curie constants according to its orientation in the magnetic field. This suggests very forcibly that whatever these characteristic quantities really stand for, it depends in some way upon the usual (most stable) orientations of the unitary magnet within the crystal. These, in turn, must be fixed by the mutual effect of adjacent atoms in that systematic arrangement which distinguishes crystals from fluids. As regards paramagnetic, as distinct from ferromagnetic, crystals, theory has made but little progress in the analysis of these effects.

Weiss, who explains the behavior of ferromagnetics by postulating a molecular field proportional to the magnetization, must, in paramagnetics which do not obey Curie's unmodified law, suppose a molecular field of another sort, independent of temperature and magnetization, which takes charge above the Curie point.[50]

Forrest, whose results for diamagnetic copper crystals have already been mentioned,[31] also studied a few paramagnetic crystals, some cubic and some not. He uses his data to determine the space-lattice according to which the unitary magnets must be arranged if the theory proposed for ferromagnetic crystals by Peddie,[51] and extended by himself, is here applicable. This amounts to regarding paramagnetism as a residual effect of ferromagnetism at temperatures where the latter is inappreciable by direct means.

The gyromagnetic ratio for a paramagnetic material—dysprosium oxide—has recently been measured by Sucksmith;[52] all previous values depend upon experiments with ferromagnetic materials. The agreement of the new value with the old indicates that the experiments give information about the unitary magnets upon the existence of which paramagnetic properties depend rather than upon their particular mode of association in ferromagnetics. All the recent data may therefore be discussed with propriety in this section.

The gyromagnetic ratio, *i.e.* the ratio of the angular momentum of a Weber element to its magnetic moment, was predicted by Richardson[53] on the assumption that the Weber element was composed of one or more electrons moving in orbits. If we use $e/mc = 1.761 \cdot (10)^7$, as recommended by Birge[5] for intra-atomic calculations, Richardson's value, $2mc/e = 1.136 \cdot (10)^{-7}$.

[50] For details of Weiss' theories see especially: Le Magnétisme by P. Weiss and G. Foëx, Colin, Paris, 1926.

[51] W. Peddie, Proc. Roy. Soc. Edinb. **25**, 1025–1059 (1905); **28**, 643–651 (1908); **32**, 216–246 (1912); **47**, 165–176 (1927); Nature **120**, 80–81 (1927).

[52] W. Sucksmith, Proc. Roy. Soc. [A] **128**, 276–293 (1930).

[53] O. W. Richardson, Phys. Rev. **26**, 248–253 (1908).

The best experimental values[54] are all about half as great. Barnett's experiments yield consistently 1.05 mc/e with an assigned probable error of about one percent. The English investigators, who put their probable error even lower, get 1.00 mc/e. No explanation of this discrepancy can yet be suggested.

As paramagnetic effects must cease to be apparent when all the unitary magnets are co-directed, while diamagnetic effects depend directly upon the applied field intensity up to the maximum intensities yet available, it is apparent that at a sufficiently high field intensity all paramagnetics should become diamagnetic. Such transitions have, indeed, been observed[55] in the case of alloys which are only feebly paramagnetic at best.

Ferromagnetic Crystals

A long series of investigators has attempted to explain the stability of magnetization which distinguishes ferromagnetic media from all others, by invoking the mutual action of unitary magnets in crystalline array. This attempt has become especially important since x-ray studies have shown that all ferromagnetic media are predominantly crystalline.

The distinguished leader in the attack upon this fundamental problem of ferromagnetism is Sir [J.] Alfred Ewing. He has had the curious experience of having a theory, which he himself abandoned as inadequate, adopted, extended and vigorously supported by others.[56] It will form a convenient thread of continuity in our discussion of recent experiments and their significance, to follow in detail the advances and retreats along this section of our battle-front.

Ewing

In its original form[57] Ewing's theory of ferromagnetism postulated an extended group of identical molecular magnets, not magnetic dipoles but of linear dimensions comparable with intermolecular distances, and made no hypothesis regarding the structure of these magnets. We may now replace "molecule" by "atom" with some gain in precision, since we now think that we know (opine?) that the atoms in—for example—iron crystals are not associated into distinguishable molecules. It then appeared, after a very simple analysis, that such magnets, centered at all the corners of a square-meshed plane net with all their axes parallel and in the plane of the net, are especially stable against disturbance by an applied magnetic field when their

[54] S. J. Barnett, Phys. Rev. [2] **6,** 239–270 (1915); **10,** 7–21 (1917); **30,** 964–965 (1927); **31,** 1116 (1928); **36,** 789 (1930); S. J. Barnett, L. J. H. Barnett, Proc. Amer. Acad. **60,** 127–216 (1925); A. P. Chattock, L. F. Bates, Phil. Trans. Roy. Soc. [A] **223,** 257–288 (1922); W. Sucksmith, L. F. Bates, Proc. Roy. Soc. [A] **104,** 499–511 (1923); W. Sucksmith, Proc. Roy. Soc. [A] **108,** 638–642 (1925).

[55] K. Overbeck, Ann. d. Physik [4] **46,** 677–697 (1915).

[56] It is interesting to notice that the whole Japanese school of magneticians, some of whom are here referred to, derived their inspiration from Ewing and Knott during the stay of these Scotch physicists at the University of Tokio.

[57] J A. Ewing, Proc. Roy. Soc. **48,** 342–358 (1890); Phil. Mag. [5] **30,** 205–222 (1890).

parallel magnetic axes are directed along one of the two sets of lines forming the sides of the squares.

As a randomly directed magnetic field in the plane of the net is increased from zero to a sufficiently high value—the initially chosen direction of this field remaining ulaltered—reduced to zero and increased in the opposite sense, the predicted behavior of the magnetic moment of the aggregate, or magnetic complex, is strikingly similar to the observed behavior of the magnetic moment of a real piece of iron held stationary and subjected to the same treatment. The more gradual changes of I_p, the component of I parallel to the applied H, H_a, in a real specimen as compared with the abrupt changes in the model were naturally ascribed to the presence in the former of a multitude of magnetic complexes with their easy directions of magnetization pointing in every possible direction with respect to the single direction of H_a.

To build up a three-dimensional model Ewing supposed the square-meshed nets to be superposed as necessary. He did not re-analyze the problem, however, because with the given assumptions it appeared that each atomic magnet is controlled almost wholly by its two nearest neighbors in the row along which it points when H_a is zero.

The models which Ewing[58] and others[59] have constructed for demonstration purposes exhibit the principal phenomena of ferromagnetism in a highly satisfactory manner. Attempts at fixing the length of an atomic magnet lead, on the other hand, to serious difficulties. We now know accurately enough the atomic spacing and arrangement in every crystal known to be ferromagnetic. We also know from saturation data the average magnetic moment per atom, and except for a few speculations,[60] it is universal practice to assume that each atom has this magnetic moment, rather than that some have more and some less. This leaves us only the ratio of the length of the magnet to the lattice constant at disposal in fixing the transverse field intensity at which the initial alignment is completely upset, and in fixing the corresponding degree of measurable magnetization, I_p. Ewing himself was the first to recognize that no single value of this ratio would satisfy the experimental data. If the ratio is made nearly as great as unity the maximum stable deflection may be made as small as the narrow range of "elastic" magnetization requires. (The refinement of measurement continually narrows this range.) But then the intensity of applied field necessary to upset the initial arrangement becomes much too great. Ewing finally[61] avoided the difficulty by shortening the magnet responsible for I and introducing for its proper control a symmetrical group of magnets fixed in the atom, and presumably also fixed with respect to the crystal space-lattice. The analogy of

[58] J. A. Ewing, Proc. Roy. Soc. Edinb. **47**, 141 (1927).

[59] Most recently: W. M. Roberds, Rev. Scient. Instr. [2] **1**, 400–403 (1930).

[60] P. Weiss, J. de phys. [6] **5**, 129–152 (1924); E. C. Stoner, Magnetism and Atomic Structure, Dutton, New York, 1926, pp. 171–172, 339; R. Forrer, Comptes rendus **182**, 1272–1275, 1530–1532; **183**, 121–123, 559–560, 738–741 (1926).

[61] J. A. Ewing, Proc. Roy. Soc. Edinb. **42**, 97–128 (1922); less completely: Proc. Roy. Soc. [A] **100**, 449–460 (1922); Phil. Mag. [6] **43**, 493–503 (1922).

this new model with the atomic models of even date was properly emphasized and it was shown that Weber elements could be so arranged as to meet the requirements, thus reducing the artificiality of the picture in terms of bar magnets.

Hysteresis energy losses, on Ewing's theory in either of its forms, must be accounted for by dissipation of magnetic potential energy whenever sudden changes of the orientation of atomic magnets occur. Transient eddy currents and magnetostrictive shocks are two possible modes of transfer of this energy from an atom to its environment during and after such changes.

Honda and Okubo

Between the first and second editions of Ewing's theory, Honda and Okubo[62] had developed his first ideas in somewhat greater detail, and still claim[63] to avoid the objection later raised by their originator. They make the atomic magnets much shorter than the least distance between atom centers and thus relatively enhance the influence of non-adjacent atoms. They also consider, as Ewing does not, the effect of neighboring complexes with their elements in various orientations, and conclude that the intensity and direction of the resultant applied field within a single complex may be very different from $\boldsymbol{H}_a$ due to external agencies. With this proviso it is easy to see that even a very small $\boldsymbol{H}_a$ may combine with the local field in such a way as to make some of the complexes give way. Hysteresis may thus become apparent at low values of H_a, and of *average* I_p, as experimental data require. Honda and Okubo, like Ewing, confine their analysis to a regular arrangement of atoms lying in a single plane, and suffering deflections only therein.

Peddie

The earliest attack upon the general problem of ferromagnetism in three dimensions was launched by Peddie.[51] He did not, however, consider a step-by-step rotation of atomic magnets in oblique or transverse fields but only the magnetic stability of co-directed magnetic doublets arranged on a space-lattice. His method was to calculate the magnetic field at a lattice-point due to equal doublets at all the other lattice-points of an infinite lattice. If this field coincides in direction with the direction chosen for the doublet axes the arrangement is inherently stable and the degree of stability is measured by the computed field intensity. If the computed field direction is not that of the doublet axes an applied field with a suitable component perpendicular to the doublet axes is necessary and sufficient for stability. Under the assumptions all changes of orientation of the atomic magnets are continuous—this is a consequence of the assumption of doublets—so that Peddie's theory makes no provision for that magnetic hysteresis which is the distinguishing feature of ferromagnetism. What it is well able to deal with is the difference

[62] K. Honda, J. Okubo, Sci. Rep. Tohoku Imp. Univ. [I] **5**, 153–214 (1916); **6**, 183–196 (1917); Phys. Rev. [2] **10**, 705–742 (1917); **13**, 6–26 (1919).

[63] K. Honda, Dictionary of Applied Physics **2**, 516–526 (1922); K. Honda, J. Okubo, Sci. Rep. Tohoku Imp. Univ. [I] **12**, 27–32 (1923).

in direction of I and H_a which is observed in all ferromagnetic crystals except for a few directions of H_a. Since the theoretical I is constant in magnitude, and always has the value corresponding to magnetic saturation, Peddie's theory is also inapplicable to low values of H_a, for which it is observed that I has smaller values.

Forrest[31] extended this theory to crystals of the orthorhombic system and applied it, as has already been stated, to both diamagnetic and paramagnetic crystals. He can only obtain qualitative results because, as he admits, thermal agitation must, at any instant, effectively free many atoms from the magnetic control which would be due to a perfectly orderly array. That is, there is no approach to magnetic saturation. His results are peculiar in that the space-lattice indicated by this partial application of theory to his measurements, or to those of others, is often not the space-lattice disclosed by x-rays.

Gans

A strikingly different point of view is that taken by Gans.[64] Instead of assuming rigid parallelism of the unitary magnets within the region considered he assumes their axes arranged in space according to a general distribution function which can be specialized to suit various initial conditions. The direction of that part of the local field intensity at any magnet which depends upon the positions and orientations of the others—what Gans calls the magnetic molecular field—is itself dependent upon the magnetic molecular field values at these other magnets, so that the general solution for I in terms of H_a proves unmanageable.[65] Special solutions are accordingly worked out for vanishing H_a and correspondingly random orientation of the magnets, and for infinitely great H_a and corresponding parallelism of all the magnets. Surprizingly enough these special solutions extrapolated respectively to high and low H_a agree so well over a wide range of I_p that Gans concludes that either may be used without sensible error. If this be admitted the study of magnetization curves can give little or no information concerning the initial distribution of unitary magnets having only magnetic constraints. It should be emphasized in this connection that this first theory of Gans ignores the crystalline character of the medium by postulating that the magnetization shall always be parallel to the applied field, and that it equally ignores hysteresis.

Much more recently, Gans[66] has proposed a theory for the magnetization of polycrystalline metal which includes more of the essential phenomena. Each crystallite is supposed to have a single direction of easiest magnetization and a sensibly rectangular hysteresis loop. The magnetization curves and hysteresis loops computed on assuming further that the distribution of

[64] R. Gans, Gött. Nachr., 118–164 (1910); 197–273 (1911); Ann. d. Physik [4] **63**, 382–396 (1920).

[65] The magnetic molecular field for a paramagnetic gas has also been considered by Gans—Ann. d. Physik [4] **50**, 163–198 (1916)—with somewhat better success.

[66] R. Gans, Schriften d. Königsb. gelehrten Ges. Naturwiss. Kl. **6**. 2, 79 pp. (1929). A long abstract is given by W. Steinhaus, Phys. Ber. **11**, 768–769 (1930).

unique directions of magnetization in space is at random agree very well with curves regarded as typical of well annealed metal. In this computation the effective field within each crystallite is divided into three parts: the applied field, the field due to the combined action of remote parts of the specimen, and the field due to adjacent crystallites. The direction and magnitude of the third component alone varies from point to point so as to account for the different values of applied field and of average magnetization at which discontinuities occur.

On the experimental side Gans[67] has succeeded in demonstrating the underlying diamagnetism of a ferromagnetic alloy—permalloy—in a field exceeding 8000 gauss, say 150000 times greater than H_c in this case.

Mahajani

A new start at a general theory of ferromagnetism has been made by Mahajani.[68] Like all the theorists yet mentioned in this section, he makes the stabilizing mechanism purely magnetic. Like Ewing, he considers the unitary magnets to have linear dimensions comparable with inter-atomic distances. In accordance with previous practice, too, he defers consideration of temperature effects. In the quantitative part of his work Mahajani uses the experimental results of Webster[69] on iron and of Weiss[70] on pyrrhotite.

The only improbable assumption in this theory—aside from the perfection of the crystal—is that the atomic magnets associated with one of the two simple cubic space-lattices into which the body-centered cubic space-lattice of iron may be decomposed, are always all parallel to one line, those associated with the other simple cubic lattice, to another. Since the points of these two lattices ought to be crystallographically equivalent there is no excuse for this postulate on physical or chemical grounds. Fortunately the quantitative successes of the theory, which belong exclusively to nearly saturated conditions, do not depend upon this assumption, the energies being computed in a straightforward way for the complete space-lattice.

Since Mahajani compares his theory with experiment by computing a "molecular field" and observing its agreement in magnitude with the quantity of this name computed from experimental data it will be necessary to explain what is meant by the term, which is due to Weiss.[71] The molecular field, $\boldsymbol{H}_m$, is a vector, having the dimensions of magnetic field intensity but regarded by Weiss as not arising from magnets, which is, in an isotropic medium, parallel to the magnetization, $\boldsymbol{I}$, and proportional to its magnitude, I, i.e. $H_m = m \cdot I$. In a crystal it is not in general parallel to $\boldsymbol{I}$, but its components $(H_m)_x$, $(H_m)_y$, $(H_m)_z$, referred to principal axes of magnetization—

[67] R. Gans, Naturwiss. **18**, 184–185 (1930).

[68] G. S. Mahajani, Proc. Camb. Phil. Soc. **23**, 136–143 (1926); Phil. Trans. Roy. Soc. [A] **228**, 63–115 (1929).

[69] W. L. Webster, Proc. Roy. Soc. [A] **107**, 496–509 (1925).

[70] P. Weiss, Arch. des sci. [4] **19**, 537–558; **20**, 213–230 (1905); J. de phys. [4] **4**, 469–508, 829–846 (1905).

[71] P. Weiss, J. de phys. [4] **6**, 661–690 (1907); Arch. des sci. [4] **37**, 105–116, 201–213 (1914).

whereon $\boldsymbol{H}$ and $\boldsymbol{I}$ are parallel—are put by Weiss proportional to the components of I along these axes:

$$(H_m)_x = m_x \cdot I_x,$$
$$(H_m)_y = m_y \cdot I_y,$$
$$(H_m)_z = m_z \cdot I_z.$$

It is of course the torque $\boldsymbol{H}_m \times \boldsymbol{I}$ which balances the torque $\boldsymbol{H}_a \times \boldsymbol{I}$, and thus accounts for the angular separation of $\boldsymbol{I}$ from $\boldsymbol{H}_a$. In a cubic crystal we would expect $m_x = m_y = m_z$, but this brings us back to Thomson's prediction[30] of magnetic isotropy in such media. Ferromagnetic cubic crystals are certainly not isotropic so we must, as Webster[69] pointed out, abandon the simple equations for $(H_m)_x$, $(H_m)_y$, $(H_m)_z$, replacing them by more complicated forms. Mahajani shows that the requirements of symmetry are met by the expressions:

$$(H_m)_x = m_x(a_1 I_x + a_3 I_x^3 + a_5 I_x^5 + \cdots),$$
$$(H_m)_y = m_y(a_1 I_y + a_3 I_y^3 + a_5 I_y^5 + \cdots),$$
$$(H_m)_z = m_z(a_1 I_z + a_3 I_z^3 + a_5 I_z^5 + \cdots).$$

For cubic crystals two terms must be considered; for non-cubic crystals one term is enough. The quantities derived from experiment are the maxima of the molecular field, i.e. $(H_m)_x$ for I along the x-axis, etc. In Webster's experiments these are of the order of 500 gauss. To fit this Mahajani must assume the radius of his Weber element to be about $2 \cdot (10)^{-9}$ cm, a not unreasonable value for an electronic orbit.

Akulov

The novelty of Akulov's treatment[72] of ferromagnetic crystals lies in the fact that he considers not only the torque exerted upon each atomic magnet by the whole array, but also the anisotropic cohesive forces which arise when the magnets are co-directed in various crystallographic directions. For the computation of these latter it is sufficient to regard the units as magnetic dipoles. Akulov is thus enabled to account qualitatively for the maximal magnetostrictions observed in iron by Webster[73] and by Honda and Mashiyama.[74] These changes in dimensions, unlike the saturation values of magnetization, are entirely different for different crystallographic directions, and therefore afford critical means for testing the theory. The longitudinal strains predicted by Akulov's theory are just about half as great as those observed by Honda and Mashiyama for the various possible directions in $\{100\}$ planes. In other planes the predicted and observed values cannot be made to agree throughout by merely introducing a constant factor, but the variations of strain with direction of applied field are qualitatively the same.

[72] N. Akulov, Zeits. f. Physik **52**, 389–405 (1928); **54**, 582–587; **57**, 249–256 (1929); **59**, 254–264 (1930).

[73] W. L. Webster, Proc. Roy. Soc. [A] **109**, 570–584 (1925).

[74] K. Honda, Y. Mashiyama, Sci. Rep. Tohoku Imp. Univ. [I] **15**, 755–776 (1926).

In order to explain non-parallelism of I and H_a in non-saturated states, Akulov, like all before him, has to extend the unitary magnet in space so that it acquires a quadrupole moment as well as a dipole moment. He notes, however, that this quadrupole moment may depend upon structural details not determined solely by the Weber element itself. In testing his theory Akulov uses the magnetization curves giving I_p as a function of H_a for single iron crystals as plotted by Honda and Kaya.[75] These curves require correction because the self-demagnetizing field varies in a more complicated way than that assumed by these investigators.[76] Since Akulov's theory ignores hysteresis it can only apply with rigor to relatively intense magnetizations where the width of the hysteresis loop becomes negligible. The mode of comparison between theory and experiment is a little difficult to describe. The area between the I_p-axis and the I_p-*vs.*-H_a curve is measured up to a line drawn parallel to the H_a-axis at a distance I_s therefrom. This area changes as H_a swings about in a chosen crystallographic plane containing two axes for which I is parallel to H_a. The curve connecting the area thus measured with the azimuth of H_a is closely reproduced by the theory. The agreement becomes more striking when it is remembered that only one parameter, the quadrupole moment, is at disposal.

The magnitude of the transverse component of magnetization, I_n, furnishes a second value of the quadrupole moment. The two values are indistinguishable and their mean is $4.52 \cdot (10)^{-29}$. The axis of the quadrupole moment coincides with that of the dipole moment. As the value for the dipole moment is about 10 Weiss magnetons, the linear dimensions necessary to explain the quadrupole moment are of the right order of magnitude.

Nickel crystals are considered in the fourth paper of this series. The theoretical predictions are compared with the data of Mashiyama[77] on magnetostriction and with those of Kaya,[78] and of Sucksmith, Potter and Broadway[79] on the relations of I_p and I_n to H_a. It is not clear from the statements here made whether or not a numerical factor has been applied to bring theory and experiment into quantitative agreement—as was necessary in the case of iron crystals. The qualitative agreement, at least, is astonishingly good.

In extending his theory to lower field intensities Akulov assumes, contrary to custom in this subject but with fair plausibility, that the atomic magnets point in all possible directions when the crystal is unmagnetized, that is that their orientations are in statistical disorder. He also assumes spatial quantization with respect to H_a. The local magnetic field arising from neighboring atoms is thus ignored. Whether this omission vitiates the result is doubtful in view of the much more serious restriction he introduces in supposing that there are only two stable orientations for each magnet, one parallel to H_a, the other anti-parallel. The relations which hold in ferromagnetic

[75] K. Honda, S. Kaya, Sci. Rep. Tohoku Imp. Univ. [I] **15**, 721–753 (1926).
[76] D. D. Foster, R. M. Bozorth, Nature **125**, 525 (1930).
[77] Y. Mashiyama, Sci. Rep. Tohoku Imp. Univ. [I] **17**, 945–961 (1928).
[78] S. Kaya, Sci. Rep. Tohoku Imp. Univ. [I] **17**, 639–664 (1928).
[79] W. Sucksmith, H. H. Potter, L. Broadway, Proc. Roy. Soc. [A] **117**, 471–485 (1928).

metals between magneto-resistance, elasto-resistance and magnetostriction seem inconsistent with a simple two-way magnetization in the general case.[80]

Starting from these more or less debatable postulates Akulov arrives at the following results for magnetization. If we take areas between I_p-*vs.*-H_a curves not up to I_s as before, but up to any arbitrary value I_p, we find the curves $\int_0^{I_p} H \, dI_p$-*vs.*-ϕ, where ϕ is the azimuth of H_a are derivable from the curve $\int_0^{I_s} H \, dI_p$-*vs.*-ϕ by a mere change of scale. If we keep I constant in magnitude by a suitable choice of H_a as the direction of I is allowed to vary we find that $|H_a \times I|$ is the same function of the direction of I for all values of I. Both of these predictions are verified by the experiments which are quoted. A third consequence of the theory is a formula by which may be calculated the difference $H_a - (H_a)_0$, where H_a and $(H_a)_0$ are, respectively, the magnitude of the applied field in a random direction, and that in a principal direction of magnetization, corresponding to a given value of I_p. (In the case of $(H_a)_0$, $I_p = I$, in the case of H_a, $I_p < I$.) This relation also appears to be supported by experiment.

As regards magneto-resistance Akulov assumes that the proportional change in resistance for a current flowing along a single-crystal wire when an intense transverse magnetic field is applied can be derived from the corresponding magnetostrictive strain by using a suitable numerical factor—of the order of 10^3—which is the same for all transverse directions of H_a in a given wire and for wires with axes in all possible crystallographic directions. The data of Kaya[81] support this generalization. Akulov also offers some evidence for supposing that the numerical factor just mentioned is invariant as we pass from saturated to non-saturated states by decreasing H_a, and that a relation of equal simplicity applies to longitudinal magneto-resistance and magnetostriction effects, at least over the range in which the effects do not show hysteresis with respect to the applied field.

All in all, the successes of the theory are impressive and it represents a notable advance in dealing with that considerable body of ferromagnetic phenomena which is most important in magnetic fields of an intensity notably greater than the coercive force, H_c. Here it is not unreasonable, as a first approximation, to suppose that the whole array of atomic magnets wheels together under the action of an increasing magnetic field constant in direction. In much weaker fields the behavior of ferromagnetic media is almost certainly more complicated than can be accounted for by Akulov's theory.

Becker

Becker[82] has recently re-emphasized the fact that a cubic arrangement of dipoles has no magnetic stability when co-directed.[83] Some further deviation

[80] For a discussion of these relations, here neglected, see L. W. McKeehan, Phys. Rev. [2] **36**, 948–977 (1930).

[81] S. Kaya, Sci. Rep. Tohoku Imp. Univ. [I] **17**, 1027–1037 (1928).

[82] R. Becker, Zeits. f. Physik **62**, 253–269 (1930).

[83] E. Ising—Zeits. f. Physik **31**, 253–258 (1925)—showed that this is true for even a single row of magnets.

from cubic symmetry—beyond that involved in the assignment of a common direction for the dipole axes—must be introduced. Two possibilities are considered. One is that already laid at the foundation of Ewing's theory and employed in slightly different forms by Mahajani and Akulov. This is the assignment of suitable linear dimensions—with corresponding quadrupole moment—to the unitary magnets. The other possibility lies in a proper distortion of the space-lattice, and this is preferred by Becker for two reasons. It permits a closer correlation of magnetic properties with mechanical strains and it avoids the inconsistency which Mahajani and Akulov fall into when they give up spontaneous magnetization in zero field and still ignore hysteresis losses in single crystals.

The theory of the distorted dipole lattice is of particular interest to those who have been driven to conclude that mechanical constraints are the chief, if not the only, source of ferromagnetic stability and of the resulting hysteresis.

For simplicity, but admittedly only as a tentative approximation, Becker assumes that I lies always in the plane determined by H_a and the strain-determined spontaneous magnetization, I_0. The form and dimensions of the regions within which I_0 may be considered constant in direction are left open for speculation. If the regions contain but a few atoms each we have much the same picture as the one I drew in a previous paper,[84] in which the stable orientation of each atom was fixed mechanically by the distances and orientations of its nearest neighbors. (This hypothesis has the awkward consequence that by making it we lose the convenience of applying the results of lattice theory, which assumes identical elements at every point of an indefinitely extended space-lattice.)

Becker's result for the longitudinal magnetostrictive strain along a four-fold axis in an iron crystal comes out about one-fifth of the observed value, or less than half of that computed by Akulov. This discrepancy must indicate a mistake in one or both of the computations, because their postulates, and the data used for comparison, are all identical. In nickel crystals the case is even worse as the predicted effect is of the wrong sign. (Akulov merely changes the sign of a coefficient to avoid mentioning this difficulty.)

A magnetization curve appropriate to polycrystalline metal is derived by Becker on the assumption that the local strains, which fix in each region the direction of its initial magnetization, are large in comparison with magnetostrictive strains—a highly reasonable hypothesis—so that the change in the elastic energy due to rotating the magnets in any region may be negligible. The degree of strain in each region, and the direction of magnetization which the strain stabilizes, together fix a critical field intensity, different for different directions of H_a, at which discontinuous changes in I_p take place for sufficiently wide ranges in H_a. There thus appear a coercive force, a remanent magnetization, a single Barkhausen discontinuity, and a hysteresis energy loss per cycle, for each region. A proper averaging over all the regions leads

[84] L. W. McKeehan, Phys. Rev. [2] **25**, 274–279 (1925).

to the sort of hysteresis loop usually found. The initial susceptibility can be computed on the same basis.

It is estimated that the maximum H_c explicable in this way by purely elastic strains is 140 gauss, a good approximation to the facts of experience. The coercive force, in Becker's theory, depends directly upon the numerical factor which fixes the magnitude (and sign) of the longitudinal magnetostrictive strain, so that he finds his theory substantiated by the low value of H_c in permalloy.[85] The wide variation of remanent magnetization, I_r, in different materials is accounted for by wide differences in the distribution function for the directions of I_0 in the unmagnetized specimen.

There is, of course, abundant evidence for the intimate relation of ferromagnetic and mechanical properties. Among recent experiments which point to such a relation are those by Kussmann and Scharnow[86] on solid solutions, where hardness and H_c increase together. It may also be noted that in the hydrogenized iron of Cioffi,[87] the low coercive force is accompanied by abnormal mechanical softness.

Elastic stress on the specimen as a whole will be expected on Becker's theory to alter the coercive force and remanence in the general way in which they have been observed to vary. It is interesting to notice in this connection that Ewing[57] stated very clearly the importance of elastic strain in changing the equilibrium orientation of the magnets in his original model. He did not, to be sure, imagine an initial strain, which is the real contribution made by Becker to the general solution.

Raising the temperature should also have a predictable effect upon H_c which should diminish in the same ratio as I_s. Less rapid variation is reported by Forrer.[48]

A prediction of Becker's theory which has not yet been tested is that negative Barkhausen discontinuities may occur in a special case—that of nickel in tension—perhaps accessible to study.

It will be observed that Becker's ideas regarding the uniqueness of the direction of easy magnetization in the homogeneous regions, and regarding the dependence of gross properties upon the distribution function for these directions, are similar to those recently offered by Gans,[66] to which reference has already been made. The superiority of Becker's presentation, as compared with that of Gans, lies in the opportunity it affords for computing the coercive force without recourse to magnetic data.

Experiments on the Barkhausen effect should properly be considered in their relation to Becker's theory. It must suffice, however, to mention some of the more striking results, leaving it to those working in this field to recognize the obvious connections and subject them to quantitative tests. The

[85] L. W. McKeehan, P. P. Cioffi, Phys. Rev. [2] **28,** 146–157 (1926).

[86] A. Kussmann, B. Scharnow, Zeits. f. Physik **54,** 1–15 (1929); Zeits. f. anorg. Chem. **178,** 317–324 (1929).

[87] P. P. Cioffi, Nature **126,** 200–201 (1930).

most quantitative experiments yet to hand in this field are those of Bozorth,[88] whose analysis of the high-frequency output of search coils surrounding a wire subjected to a smoothly varying magnetic field leads to a value for the mean change in magnetization in a Barkhausen discontinuity at various points on the hysteresis loop, and suggests that discontinuities account for most if not all of the whole change in magnetization where loops are steep. It is not yet clear whether the 10^{15} or so atoms which must cooperate in a discontinuity of average size are all the atoms in a sharply bounded region, or whether they form chains, or network, with unaffected atoms lying between them. The smallness of the effect of grain-size upon ferromagnetic properties,[89] and the fact that Barkhausen discontinuities of ordinary size are found to occur in large mono-crystals, make it clear that the grain or crystallite of the metallographist is not magnetically homogeneous. Only when a substance is finely divided,[90] or made into a very thin film,[91] do its intrinsic magnetic properties suffer any important changes.

Other notable papers on the Barkhausen effect are those by Heaps and his students[92] which deal with the effects in minute specimens, where the discontinuities may be few and relatively very large. The most recent results show that large discontinuities in magnetization are accompanied by discontinuous changes in length, *i.e.* by magnetostrictive jumps. Such changes in length had previously been observed by Arkadiew[93] who simply listened to a telephone receiver diaphragm actuated by the tightly stretched specimen as it was carried through a magnetic cycle, but the simultaneity of length and magnetization changes was not then demonstrated. Studies by Preisach[94] show that the jumps in I_p may be made enormous (or run together so as to be difficult of separate observation) by tension or torsion, and Sixtus and Tonks[95] report a kind of trigger action by which a local increase in H_a can set off a discontinuity which progresses far beyond the region in which it started. The relatively slow propagation of such waves of magnetization—much slower than the velocity of sound in the medium—may be wholly accounted for by eddy currents. It is attractive to speculate, however, that

[88] R. M. Bozorth, Phys. Rev. [2] **34,** 772–784 (1929); R. M. Bozorth, J. Dillinger, Phys. Rev. [2] **35,** 733–752 (1930).

[89] O. v. Auwers, Wiss. Veröffentl a. d. Siemens-Konz. **4,** 266–273 (1925); **7,** 197–209 (1928); Zeits. f. techn. Physik **6,** 578–582 (1925); **9,** 475–478 (1928); K. Daeves, Zeits. f. techn. Physik **10,** 67–68 (1929); O. v. Auwers, Zeits. f. techn. Physik **10,** 68 (1929); G. J. Sizoo, Zeits. f. Physik **51,** 557–564 (1928); **53,** 449–457; **55,** 144 (1929); O. v. Auwers, G. J. Sizoo, Zeits. f. Physik **60,** 576–580 (1930); T. D. Yensen, Metals & Alloys **1,** 493–495 (1930).

[90] L. A. Welo, O. Baudisch, Phil. Mag. [7] **3,** 396–410 (1927); W. Köster, Zeits. f. anorg. Chem. **179,** 297–298 (1929).

[91] The ferromagnetic properties of thin films are still very puzzling and the conclusions that can as yet be drawn are, with few exceptions, empirical. For a recent contribution to one phase of the subject see J. H. Howey, Phys. Rev. [2] **34,** 1440–1447 (1929).

[92] C. W. Heaps, J. Taylor, Phys. Rev. [2] **34,** 937–944 (1929); C. W. Heaps, A. B. Bryan, Phys. Rev. [2] **36,** 326–332 (1930).

[93] W. Arkadiew, Comptes rendus **184,** 1233–1234 (1927).

[94] F. Preisach, Ann. d. Physik [5] **3,** 737–799 (1929).

[95] K. J. Sixtus, L. Tonks, Phys. Rev. [2] **35,** 1441 (1930).

the delay may be due in part to the necessity for waiting before each advance for a favorable condition as regards energy of thermal agitation, and that we may yet trace a continuous progression of phenomena from the almost instantaneous response of the atomic magnet to the slow secular variation of magnetic moment in well-aged magnet steel. If so, magnetic lag—the reality of which has been a matter of dispute for forty years[96]—may be admitted as an intrinsic ferromagnetic phenomenon.

The enlargement or coalescence of microscopic Barkhausen discontinuities into major catastrophes in magnetization was first observed by Forrer[97] in suitably over-strained nickel wires. The abnormal values of remanence which he observed led him to postulate a rather complicated unitary magnet in this metal, the two parts of which might take up either of two configurations in the absence of applied field. The high remanence corresponded to a preponderance of the usually less numerous configuration. He extended the same reasoning to iron, in which he thinks that the unit must have three parts and in a recent paper[98] offers different types of hysteresis loop in iron as experimental evidence in favor of his hypothesis. This effect may, of course, be explained in other ways. The dependence of remanence upon the structure of the specimen is undoubtedly real and has been discussed in an earlier paragraph.

In this group of phenomena should be included the effects of applied stresses upon magnetization. These have been known from time immemorial. The most recent addition to the list is the effect of quenching stresses on he magnetization of toroids.[99] Here, too, belong any observed effects of magnetization upon mechanical hysteresis. Guye and Dupraz[100] have shown that a longitudinal magnetic field reduces the logarithmic decrement of torsional vibrations of large amplitude in wires of iron, steel and nickel. This must be taken to mean that the orientations of atomic axes are rendered more stable against discontinuous rotation under shearing stresses by the added magnetic control thus supplied.

The rate of heat production at various stages of a hysteresis cycle must be affected by the incidence and relief of the elastic strains involved in changes of magnetization, and by any mechanical hysteresis which occurs in consequence. All observers[101] agree that most of the heat is liberated in that part of the cycle in which I_p increases (or decreases) rapidly from zero. The most elaborate and careful study, by Ellwood, discloses cooling in two regions on each branch of a sufficiently extensive hysteresis loop in specimens of carbon

[96] The most recent clash: C. Lapp, Ann. de phys. [10] **8,** 278–395 (1927); **12,** 276–277 (1929); R. M. Bozorth, Phys. Rev. [2] **32,** 124–132 (1928).

[97] R. Forrer, J. de phys. [6] **7,** 109–124 (1926); **10,** 247–262 (1929).

[98] R. Forrer, J. Schneider, Comptes rendus **190,** 1391–1393 (1930).

[99] O. Tesche, Zeits. f. techn. Physik **11,** 239–242 (1930).

[100] C. E. Guye, A. Dupraz, Helv. Phys. Acta **2,** 320–342 (1929).

[101] U. Adelsberger, Ann. d. Physik [4] **83,** 184–212 (1927); F. W. Constant, Phys. Rev. [2] **32,** 486–493 (1928); K. Honda, J. Okubo, T. Hirone, Sci. Rep. Tohoku Imp. Univ. [I] **18,** 409–417 (1929); W. B. Ellwood, Phys. Rev. [2] **36,** 1066–1082 (1930).

steel. No explanation of these coolings, one of which was also observed by Adelsberger, has yet been offered.

Besides the data on single crystals made use of by Becker and his immediate predecessors we now have comparable work from other laboratories on iron[102] and nickel,[103] and new information regarding cobalt.[104] None of this material has been examined for its theoretical significance.

Heisenberg

The electron interchange interaction theory of Heisenberg[105] is a first attempt to lay down a quantum-theoretical basis for discussions of ferromagnetism. Stripped of its mathematics it seems merely to say that an atomic magnet, if it is not to be freely assignable to any of its possible magnetic states —giving rise to paramagnetism only—must have several possible states giving positive contributions to the magnetization of the medium (this requires that the quantum numbers characterizing the magnetic element must not be too small), and must have enough immediate neighbors so that its energy of interaction with these may be considerable. The electron interchange associated with the theory is not an actual transfer of an electron from one atom to another, and *vice versa*, as the word interchange might seem to imply, but only a mutual effect of two electrons, one in atom A and one in atom B, by which the states permissible to each are reduced in number, and, in particular, the state of mutual parallelism is made more probable. Seen from this stand-point we notice that we have been discussing such interchange for some time past.

As first laid down, and in its more rigorous development,[106] Heisenberg's theory gives necessary, but not necessary and sufficient, conditions for ferromagnetism. As soon as physical details begin to intrude into the mathematics, we notice an improvement in its definiteness. Stoner[107] finds it adequate to consider together in the assignment of magnetic states not all the atoms in the specimen but only a single atom and its immediate neighbors. He has gone a little too far in supposing that only two atomic orientations need to be considered, as this prevents him from considering magnetostriction phenomena[108] and ignores the crystalline symmetry of the atomic environment. Slater[109] has made important progress in discovering a reason for

[102] W. Wolman, Archiv f. Elektrot. **19,** 385–404 (1928); G. J. Sizoo, Zeits. f. Physik **56,** 649–670; **58,** 718 (1929); Physica **10,** 1–18 (1930); H. Gries, H. Esser, Archiv f. Elektrot. **22,** 145–152 (1929); Archiv f. d. Eisenhüttenwesen **2,** 749–761 (1929).

[103] G. J. Sizoo, Zeits. f. Physik **57,** 106–114 (1929); Physica **10,** 1–18 (1930).

[104] S. Kaya, Sci. Rep. Tohoku Imp. Univ. [I] **17,** 1157–1177 (1928); Z. Nishiyama, Sci. Rep. Tohoku Imp. Univ. [I] **18,** 341–357 (1929).

[105] W. Heisenberg, Zeits. f. Physik **49,** 619–636 (1928).

[106] F. Bloch, Zeits. f. Physik **61,** 206–219 (1930); E. Teller, Zeits. f. Physik **62,** 102–105 (1930).

[107] E. C. Stoner, Proc. Leeds Phil. Lit. Soc. [Sci. Sect.] **2,** 50–55, 56–60 (1930).

[108] This has been emphasized by Heaps and Bryan, *loc. cit.*[92] in criticizing Bozorth's calculation of the volume taking part in a Barkhausen discontinuity on the basis of complete reversals of magnetization.

[109] J. C. Slater, Phys. Rev. [2] **35,** 509–529; **36,** 57–64 (1930).

the unique position of iron, cobalt and nickel as ferromagnetic elements, by importing into the theory the idea that the probability of interaction increases as the ratio of interatomic distance to radius of incompleted electron shell increases. This also gives the long awaited "explanation" of the ferromagnetism of Heusler's alloys.[110] The supposition is that when copper replaced manganese in compounds of the type Mn_3X, where X is any trivalent metal, the copper must be so highly ionized that its incomplete three quantum shell, lying deep within its structure, becomes available for ferromagnetic interchange. The question which frequently arises as to the possibility of a ferromagnetic condition of manganese may receive its answer in like manner. Freese[111] has been unable to get ferromagnetic manganese by any of a variety of methods of preparation, and work by Miss Wheeler in this laboratory, as yet unpublished, has the same negative result.

The curious fact that alloys of iron and cobalt in certain proportions have a higher saturation intensity of magnetization than either element alone, discovered by Weiss,[112] has been confirmed by the systematic studies of Elmen,[113] and Masumoto.[114] The discoverer[115] has given further attention to this matter. It now appears that the maximum value of I_s lies at rather higher cobalt content than at first supposed, when it was attributed to a compound Fe_2Co, no evidence for which has yet been found by x-ray methods. The ternary alloys of iron, cobalt and nickel should offer a proving ground for theories which attribute ferromagnetic behavior to peculiarities of atomic environment. Among these alloys the group—the perminvars—discovered by Elmen[116] to have extremely constant magnetic permeability over a considerable range in H_a, with correspondingly low hysteresis losses, are especially interesting. Cobalt dissolved in a large excess of platinum displays, according to Constant,[117] a notably higher value of I_p than can be accounted for by the cobalt atoms only unless these are in an abnormal state. This may prove to be another instance of environmental influence.

The attempt by Dorfmann and his colleagues[118] to explain ferromagnetism as due to conduction electrons has met with little favor. An error in sign[119] vitiates the quantitative argument, and even qualitatively it rested solely on the fact that magnetic properties and thermoelectric properties

[110] For recent work on the crystal structures and magnetic properties of these alloys see the following: L. Harang, Zeits. f. Krist. **65,** 261–285 (1927); H. O. Doerum, Avh. Oslo 1929 No. 10, 32 pp.; E. Persson, Zeits. f. Physik **57,** 115–133 (1929); H. H. Potter, Proc. Phys. Soc. London **41,** 135–142 (1929); S. Valentiner, G. Becker, Zeits. f. Physik **57,** 283–291 (1929).

[111] H. Freese, Phys. Zeits. **29,** 191–197 (1928)

[112] P. Weiss, Trans. Faraday Soc. **8,** 149–156 (1912).

[113] G. W. Elmen, J. Franklin Inst. **207,** 583–617 (1929).

[114] H. Masumoto, Sci. Rep. Tohoku Imp. Univ. [I] **18,** 195–229 (1929).

[115] P. Weiss, R. Forrer, Ann. de phys. [10] **12,** 279–374 (1929).

[116] G. W. Elmen, J. Franklin Inst. **206,** 317–338 (1928).

[117] F. W. Constant, Phys. Rev. [2] **34,** 1217–1224 (1929).

[118] J. Dorfmann, R. Jaanus, Zeits. f. Physik **54,** 277–288 (1929); J. Dorfmann, I. Kikoin, Zeits. f. Physik **54,** 289–296 (1929); *cf.* F. Bloch, Zeits. f. Physik **57,** 545–555 (1929).

[119] E. C. Stoner, Nature **125,** 973 (1930).

change rapidly at the same temperature—no direct connection has been established. The work of Smith and Dillinger,[120] to be sure, extends the temperature range over which there is an evident relation, but it seems safer to attribute both effects to changes in structural stability rather than to make the same electrons serve in both cases. For similar reasons the Hall and allied effects seem to lie beyond the scope of this review. Attempts to localize the magnetic electrons by the diffraction of x-rays[121] have so far been fruitless.

Elastically constrained atomic magnets might be expected to show resonance by abnormal values of susceptibility at particular high frequencies. Within the range of experiments so far conducted the results have been negative.[122]

Weiss

The molecular field theory of Weiss,[70] already discussed to some extent in connection with the theory of Mahajani, is not, strictly speaking, an atomistic theory at all. It makes no use of the spatial relations of the unitary magnets. It is only mentioned again in this section because some of the recent experimental work which at present seems chiefly concerned with it will in all probability be needed shortly in a more comprehensive theory of ferromagnetic crystals.

The loss of ferromagnetic properties at a fairly definite temperature—the Curie point—is hardly explicable without resort to an atomistic theory of some sort. Recent theories which focus attention upon thermodynamic reasoning[123] are not therefore much more satisfactory in this respect than the admittedly formal equations of Weiss.[124] On the experimental side Weiss and Forrer,[125] and Tyler,[126] working with nickel, and Bates,[127] working with manganese compounds having low Curie points, have tried to make more precise the relation between I_s and T, which has a very simple form in Weiss' theory. The results are hardly consistent enough to permit any decision for or against the theoretical equation, though Weiss regards his results as conclusive in his favor. The difference between the Curie points for heating and cooling is very large in some cases—unfortunately among these is the case studied by Bates—and this difference has been shown, at least for iron-nickel alloys, to be independent of the rate of heating or cooling.[128] Until these temperature-hysteresis effects are better understood the matters at issue cannot be settled.

[120] A. W. Smith, J. Dillinger, Phys. Rev. [2] **33,** 398–402 (1929).

[121] T. D. Yensen, Phys. Rev. [2] **32,** 114–123 (1928); J. C. Stearns, Phys. Rev. [2] **35,** 1–7 (1930).

[122] G. R. Wait, F. G. Brickwedde, E. L. Hall, Phys. Rev. [2] **32,** 967–973 (1928); N. N Malov, J. Appl. Phys. Moscow-Leningrad **6,** 26–38 (1929).

[123] R. H. Fowler, P. L. Kapitza, Proc. Roy. Soc. [A] **124,** 1–15 (1929); E. C. Stoner, Phil. Mag. [7] **10,** 27–48 (1930).

[124] P. Weiss, J. de phys. [7] **1,** 1–10, 163–175 (1930).

[125] P. Weiss, R. Forrer, Ann. de phys. [10] **5,** 153–213 (1926).

[126] F. Tyler, Phil. Mag. [7] **9,** 1026–1038 (1930).

[127] L. F. Bates, Phil. Mag. [7] **6,** 593–597 (1928); **8,** 714–732 (1929).

[128] G. Gossels, Zeits. f. anorg. Chem. **182,** 19–27 (1929).

The magneto-caloric effect—the reversible rise in temperature observed when an intense magnetic field is applied—is another facet of the Curie point problem. Here all the data come from Weiss' laboratory.[129] The anomalous specific heat of ferromagnetics[130] is still another. So too is the thermo-elastic anomaly.[131]

Conclusion

This review has mentioned many as yet unsolved problems, but a glance at the too copious references with which the text has been encumbered will show how wide spread the interest in magnetism has recently become, and will encourage a hope that the present position may soon be consolidated and made a safer starting point for further advances.

It is apparent from what has been said in criticism of the various theories, particularly those which attempt to cover the behavior of ferromagnetic crystals, that a comprehensive explanation must assign the magnetic states of atomic, ionic and electronic magnets with proper regard to the local value of magnetic field intensity $\boldsymbol{H}$—not merely its applied part $\boldsymbol{H}_a$—, and that the assignment must be principally conditioned by the spatial configuration—including here even slight distortions in crystal space-lattice if the medium is crystalline—and chemical characteristics (nuclear and electronic constitution) of neighboring atoms and ions. In nearly perfect crystals at low temperatures many contiguous units may be found, to a necessary approximation, in the same magnetic state and, at least in imperfect crystals, some of the states must be extremely stable, in order to make possible the phenomena of ferromagnetism. Under paramagnetic conditions there must be equilibrium between magnetic and thermal energies. In diamagnetic media the non-decaying eddy currents must be localized and their constraints, if any, determined by their environment. The complexity of the indicated theory is such that its complete presentation may not soon occur. Meanwhile we must be careful to use the partial explanations now at hand with due caution and with frequent reference to experiment for their correction.

Sloane Physics Laboratory,
September 6, 1930.

[129] P. Weiss, J. de phys. [6] **2,** 161–182 (1921); P. Weiss, R. Forrer, *loc. cit.*[125]

[130] L. F. Bates, Proc. Roy. Soc. [A] **117,** 680–691 (1928); B. G. Whitmore, Phil. Mag [7] **7,** 125–129 (1929); E. Lapp, Ann. de phys. [10] **12,** 442–451 (1929).

[131] E. Bauer, J. de phys. [6] **10,** 345–359 (1929).

THE INTERPRETATION OF BAND SPECTRA, PARTS I, IIa, IIb. ADDITIONS AND CORRECTIONS*

By Robert S. Mulliken
University of Chicago

STRICTLY, Eqs. (3*a*), (4*a*), and (6) should read

$$B_v = B_e - \alpha_e(v+1/2) + \cdots = B_0 - \alpha v + \cdots, \text{ etc.} \qquad (3a)$$

$$D_v = D_e + \beta_e(v+1/2) + \cdots = D_0 + \beta v + \cdots, \text{ etc.} \qquad (4a)$$

$$\cdots \alpha_e = (6B_e^2/\omega_e)\ (2b^*+1) \cdots . \qquad (6)$$

Fig. 1. In Fig. 1 and elsewhere, $D_e = U(\infty) - U(r_e)$ should be distinguished from $D \equiv D_0 = U(\infty) - E(v=0)$.

P. 67. After Eq. (6), it should be noted (cf. Ref. 12*b*, also T. Hori, Zeits. f. Physik, **62,** 352, 1930) that examples where α_e and $x_e\omega_e$ are negative have now been found.

Pp. 92–95. In connection with the discussion on pp. 92–95, it should have been pointed out that the "strong field" and the "weak field" cases of Fig. 10 respectively correspond to Hund's cases *a* and *c* (in case *a*, the effect of the electric axis is strong, in case *c* it is weak, compared with the coupling between the *L* and *S* vectors).

Pp. 95, 107. On p. 95, the last sentence in the text ("The $^3\Sigma$ levels of $O_2 \cdots$") is incorrect and should be omitted. The same is true of the last sentence in the first paragraph of p. 107, beginning "(It may be recalled $\cdots$)."

P. 115. In the third from last paragraph, the recommended symbols should read "T^e or T_e" instead of "T^e," and "$\Delta G(v+1/2) = G(v+1) - G(v)$," instead of "$\Delta G(v) = G(v+1/2) - G(v-1/2)$."

— — — — — — — — — — — —

Pp. 99–100. The following changes will remove some errors, and improve the discussion in other respects:

P. 99, fifth line from bottom, replace "We have $\cdots = -2O\rho$," by" We then have $\overline{N^2} = O^2 + \overline{G^2} + \phi^r(K)/B_v$, where $\phi^r(K)/B_v = -2\overline{O \cdot G} = -2O\rho$.

P. 100, second to fifth line, delete one sentence beginning "If the rotation $\cdots$."

P. 100, replace second ("Now substituting $\cdots$") and third paragraph by the following *three* paragraphs:

The existence of a ρ implies that the motion of the L^* vector has been altered by the rotation. This change is the beginning of a transition from case *b'*, where L^* precesses around the electric axis, to case *d'*, where it precesses around the axis of rotation (cf. next section). Corresponding to the beginning "uncoupling" of L^* from the electric axis, we expect in general an alteration of F^{el}, so that we may set $F^{el} = F_0{}^{el} + \phi^{el}(K)$.

* For Parts I, IIa, IIb, see Reviews of Modern Physics **2**, 60–115 (1930).

Now substituting $O^2 = K^{*2} - \Lambda^2$ (cf. Fig. 11), and $K^{*2} = K(K+1)$, we get $\overline{N^2} = K(K+1) - \Lambda^2 + \overline{G^2} + \phi^r(K)/B_v$, and $\overline{N^4} = K^2(K+1)^2 + \cdots$. Substituting in the F^r expression, adding F^{el} and F^v, and then putting $\phi_i(K)$ in place of $\phi^r(K) + \phi^{el}(K)$, we get[38]

$$F = F_0^{el} + F^v + B_v[K(K+1) - \Lambda^2 + \overline{G^2}] + \phi_i(K) + D_v K^2(K+1)^2 + \cdots \quad (29)$$

F^v has the same meaning as previously [cf. Part I, Eq. (5)]. The small terms given by $B_v(\overline{G^2} - \Lambda^2)$ are of no practical importance, since for any electronic state and v value, they enter as additive constants, which are not separable, in the analysis of band spectra, from the large terms $F_0^{el} + F^v$.

The "uncoupling term" $\phi(K)$ in Eq. (29) has been written $\phi_i(K)$ because if $\Lambda > 0$ it is double valued [$\phi_a(K)$ and $\phi_b(K)$], as discussed in the next paragraph. If the uncoupling is small, this function takes the form $\phi(K) \sim \delta K(K+1)$, so that the coefficient of $K(K+1)$ in Eq. (29) becomes $B_v + \delta = B_v^x$ instead of B_v; usually $\delta << B_v$, but sometimes δ is large enough so that it needs to be considered if one wishes to determine B_v, B_e, and r_e accurately. If the uncoupling is larger, $\phi(K)$ takes the form $\phi(K) = \kappa + \epsilon K + \delta K(K+1) + \cdots$; when such a term is added to $B_v K(K+1)$, the sum may be expressed in the form $\kappa + B_v^x K^x(K^x+1)$, where B_v^x and K^x differ slightly from B_v and K. Thus one gets apparent K values (K^x) differing slightly from integers. Slight deviations of this sort, and probably from this cause, are often found in the analysis of band spectra.[12]

P. 100, last two lines, replace $-\delta_i$, $-\delta_a$, $-\delta_b$, by $+\delta_i$, $+\delta_a$, $+\delta_b$.

— — — — — — — — — — — — —

On p. 110 it is stated that Eq. (41), which is obtained by expanding Eq. (37) as a power series in the parameter A/B_v, represents a good approximation to Eq. (37) for a range of A/B_v values corresponding well to case b and extending, roughly, from -2 to $+6$. Eq. (41), neglecting terms in A^2, A^3, $\cdots$, is then used in obtaining Eqs. (42) and (43)–(44), which are then shown to agree with Eqs. (34) and (35). Unfortunately this procedure is unjustified except for a much more limited range of A/B_v values (say from $A/B_v = -\frac{1}{2}$ to $+\frac{1}{2}$; even here the approximation is not too good). Although the *type of coupling* of the quantum vectors is essentially that corresponding to case b for the entire range of A/B_v values from -2 to $+6$, except for the lowest J values,—and for a larger A/B_v range for high J values—the departures of the *energy* values from those of the limiting case b or b' ($A = 0$), and especially the doublet separations $F_1(K+\frac{1}{2}) - F_2(K-\frac{1}{2})$, can be correctly represented only by using the Hill and Van Vleck formula Eq. (37).

In the same way, the item (1) at the bottom of p. 106 and the terms $A\Lambda^2/K(K+1)$ in Eqs. (33) and (36) and $A\Lambda^2/2K$ or $A\Lambda^2/2(K+1)$ in Eqs. (34) and (35), are adequate only for a very limited range of A/B_v values close to $A = 0$. In other cases, these expressions should be replaced by others derived by subtracting Eq. (32), with $f(K, J-K)$ set equal to zero, from Eq. (37). Of course for Σ states ($\Lambda = 0$) Eqs. (33)–(36) are correct if we merely drop the terms in $A\Lambda^2$.

The physical reason for the invalidity of Eqs. (33)–(36) and (41)–(44) for $\Lambda > 0$ and A/B values not close to zero is as follows. Eqs. (33)–(36) make allowance for the departures from case *b* toward case *a* only by adding a *magnetic* energy term (energy of interaction of Λ and spin)—, item 1, foot of p. 106. But actually the effect of these departures on the *kinetic* energy (given in case *b* by $B_v K(K+1) + \cdots$) is usually a good deal more important. This kinetic energy correction is accurately included, together with the magnetic energy correction, in Eq. (37), but not in Eqs. (33)–(36) or (42)–(44), and these corrections are not even convergently represented by the expansion Eq. (41), except for very small A/B_v values.

Eq. (37) does not, however, include the small correction term in γ which appears in Eqs. (33)–(36) and (41)–(44). The inclusion of this term in the form given in Eq. (33) appears to be justified so long as the coupling is essentially that of case *b*. Eqs. (33) and (36) may then be rewritten

$$f(K, J-K) = \psi(A, \Lambda, K) + \tfrac{1}{2}\gamma[J(J+1) - K(K+1) - S(S+1)] + w(K, J-K) \tag{33}$$

$$\Delta f(K) = \Delta\psi(K) + \gamma(K+\tfrac{1}{2}) \tag{36}$$

where ψ is a function of A, Λ, and K which vanishes for $\Lambda = 0$ and which may be determined for $\Lambda > 0$ and $S = \frac{1}{2}$ by means of Eq. (37).

VOLUME 2 OCTOBER, 1930 NUMBER 4

REVIEWS OF MODERN PHYSICS

UNIV. OF NEBR.
NOV 1 3 1930
LIBRARY

(Formerly Physical Review Supplement)

CONDUCTED BY THE
AMERICAN PHYSICAL SOCIETY

BOARD OF EDITORS

JOHN T. TATE, University of Minnesota
A. H. COMPTON K. T. COMPTON K. K. DARROW
E. C. KEMBLE C. E. MENDENHALL D. L. WEBSTER

CONTENTS

Thermionic Emission – – – – – – – – – – – – – – – – – – SAUL DUSHMAN 381
Magnetism in Discontinuous Media – – – – – – – – L. W. MCKEEHAN 477
Interpretation of Band Spectra, Parts I, IIa, IIIb. Additions and Corrections – R. S. MULLIKEN 506

PUBLISHED QUARTERLY
BY THE AMERICAN PHYSICAL SOCIETY
PUBLICATION OFFICES: 450-454 AHNAIP STREET, MENASHA, WIS.
EDITORIAL OFFICES: MINNEAPOLIS

THE AMERICAN PHYSICAL SOCIETY

OFFICERS OF THE SOCIETY

President:	HENRY G. GALE	University of Chicago, Chicago, Ill.
Vice-President:	W. F. G. SWANN	Bartol Research Foundation, Philadelphia, Pa.
Secretary:	W. L. SEVERINGHAUS	Columbia University, New York City
Treasurer:	G. B. PEGRAM	Columbia University, New York City
Editor:	JOHN T. TATE	University of Minnesota, Minneapolis, Minn.
Local Secretary for the Pacific Coast:	L. B. LOEB	University of California, Berkeley, Calif.

REVIEWS OF MODERN PHYSICS

REVIEWS OF MODERN PHYSICS are published quarterly under the auspices of the American Physical Society. Their purpose is to provide those interested in developments in physics with comprehensive, authoritative and timely discussions of the problems which are, at the present time, of especial interest and importance.

It is intended that in so far as possible these discussions shall be written in such form as to interest and to inform those who are not specialists in the subjects under discussion.

It is hoped that the studies appearing in this journal will give to the specialist the benefit of the view point, in a broad sense, of other specialists; to the student, a background of critical knowledge and an adequate bibliography of source material; to those whose time is absorbed in the teaching of physics, a stimulating account of the forward movement of their science.

REVIEWS OF MODERN PHYSICS are published quarterly in January, April, July and October by the American Physical Society at 450 Ahnaip St., Menasha, Wis. Volume 1 comprises the issues of July and October, 1929. Volume 2 and subsequent volumes begin with the January, 1930 issue.

SUBSCRIPTION PRICE

To members of the American Physical Society or of the American Chemical Society	per year	$3.00
To all others	per year	4.00
Single copies		1.25

All editorial correspondence should be addressed to REVIEWS OF MODERN PHYSICS, University of Minnesota, Minneapolis, Minnesota.

Application has been made for entry as second class matter at the post office at Menasha, Wisconsin.

Reviews of Modern Physics

(Formerly)

The Physical Review Supplement

ARTICLES IN PREPARATION

The following reviews are being prepared and will appear in forthcoming issues of this journal.

Review of Line Spectroscopy - - - - - - - - - - - R. C. GIBBS

Illustrative description and theory of various kinds of line spectra, Landé's interval rule, Hund's rule for relative position of terms, iso-electronic sequences, regular and irregular doublet laws, series limits, ionization potentials, displacement law, intensity rules, criteria used in identification work, Zeeman effect, hyperfine structure.

Quantum Mechanics of Collision Processes - - - - - E. U. CONDON

The principles underlying and the results of the theoretical treatment of simpler problems in which particles are scattered by fields of force and by atoms.

Electrical Conductivity of Flames - - - - - - - - - H. A. WILSON

(1) Relation between current, potential and distance between electrodes in a Bunsen flame.
(2) Conductivity of alkali metal vapors in flames.
(3) Conductivity of flames for alternating currents.
(4) Mobility of the ions in flames.
(5) Effects of a magnetic field on the conductivity of flames.

Polyatomic Molecules and Their Infra-Red Spectra - DAVID M. DENNISON

In treating the problem of polyatomic molecules a review is made of the theory of their vibrations and rotations. A survey is given of the principle experimental data and these are critically discussed in the light of the theory to show in what cases the essential features of the bands may be explained.

Application of Positive Ray Analysis to Ionization Problems - H. D. SMYTH

Experimental methods, multiple ionization, products of ionization in polyatomic gases, comparison with potential energy curves and thermochemical data, collisions of the second kind, mean free paths of positive ions.

Statistical Theory of Thermoelectric, Galvano- and Thermomagnetic Phenomena in Metals - - A. SOMMERFELD AND N. H. FRANK

Quantum Mechanics of Metals - - - - - - - - - - J. C. SLATER

Electrical Discharges in Gases. Part II

- - - - - - - - - - - K. T. COMPTON AND IRVING LANGMUIR

Thermionic and Photoelectric Properties of Atomic Films - J. A. BECKER

A New Instrument

THE SOUNDMETER

For Your Acoustics Laboratory

No. 3281

Designed for the purpose of quantitatively measuring the pressure of sound waves

Sensitivity Independent of Frequency.
Approximate Sensitivity: 2 dynes per sq. cm. per scale division.

Three fundamental uses will suggest themselves for the student of acoustics:

1. **To demonstrate that sound waves produce a pressure** which can accurately be measured. This can readily be demonstrated with the setup as shown in the illustration. Either condensation or rarefaction pressure can be measured by reversing the valve.
2. **Exploration work in stationary waves.** Accessory No. 3281A is necessary and can be used as an exploring valve in a resonating pipe or organ pipe to measure the pressure at the loops and nodes.
3. **Absorption measurements may be made** of different materials placed over the end of the tube.

Complete with indicating mechanism and sound valve and a resonator tube of a length to respond to 256 v.p.s. Other accessories are also listed.

Send for more complete description and a copy of the original research report by B. E. Eisenhour and F. G. Tyzzer of Riverbank Laboratories, Geneva, Illinois.

No. 3281 Soundmeter **$50.00**
No. 3281A Probe, with valve, hose, and attachment for soundmeter **10.00**
No. 3281B Adjustable Resonator **2.50**

NEW MAGNESIUM ALLOY TUNING FORKS

Weigh about one-fourth as much as steel, and do not rust.
The time of vibration is much longer than for steel forks, or many of the alloy forks. Write for prices on all ranges.
They can be supplied, guaranteed to 0.5% accuracy.
No. 3257 Tuning Forks, magnesium, set of 3 (256, 768, and 1280 v.p.s.) **$15.00**

PACIFIC COAST REPRESENTATIVES:

Braun-Knecht-Heimann Co.
576-584 Mission Street,
San Francisco, California.

Braun Corporation,
365 New High Street,
Los Angeles, California.

W. M. Welch Manufacturing Company

Manufacturers, Importers and Exporters of Scientific Apparatus, Furniture, and School Supplies
General Offices: 1515 Sedgwick St., CHICAGO, ILL., U. S. A.

Scientific Apparatus Factory and Warehouse:
1516 Orleans St., Chicago, Ill.

Laboratory Furniture Factory:
Manitowoc, Wisconsin

Branch Offices:

342 Madison Av.
NEW YORK CITY, N.Y.

1916 West End Av.
NASHVILLE, TENN.

34th & Broadway
KANSAS CITY, MO.

2220 Guadalupe St.
AUSTIN, TEXAS

GEORGE BANTA PUBLISHING COMPANY, MENASHA, WISCONSIN

DATE DUE

Unless this book is returned on or before the last date stamped below a fine will be charged. Fairness to other borrowers makes enforcement of this rule necessary.